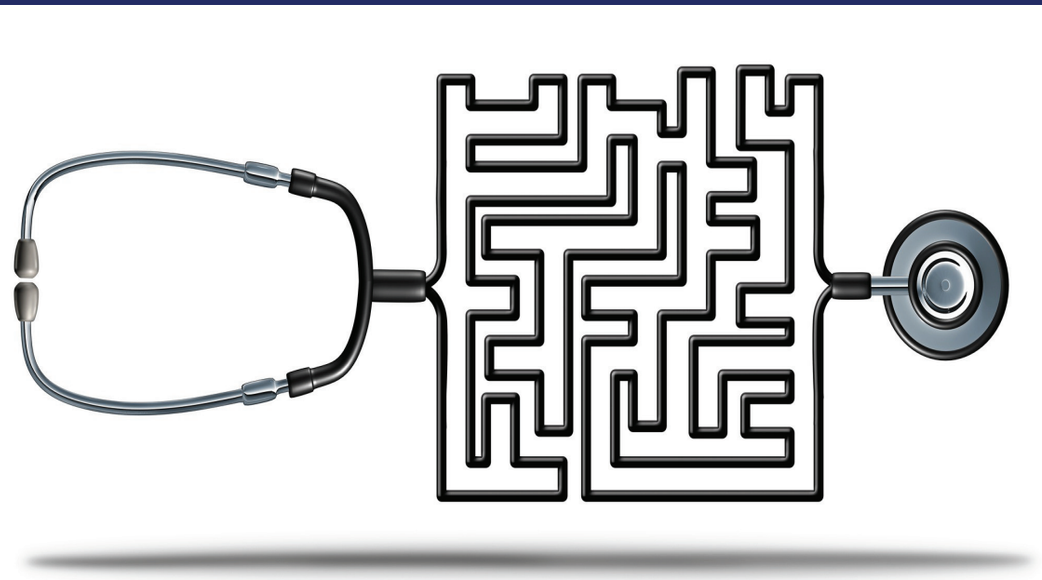
В монографии рассмотрены методы построения математических моделей для задач классификации с обучением, возникающих в медицинских приложениях. Предложен авторский метод построения моделей классификации, основанный на геометрической интерпретации структуры многомерных данных, а также методы формирования композиций классификаторов; описаны результаты их применения для решения практических задач дифференциальной диагностики, прогнозирования клинического исхода и оценки тяжести состояния пациентов. Работа будет интересна разработчикам математического обеспечения автоматизированных систем поддержки принятия решений в медицине и других областях применения прикладных методов математического моделирования; научным работникам; врачам; студентам и аспирантам технических и медицинских специальностей.

Нессонова Марина Николаевна, кандидат технических наук по специальности "медицинская и биологическая информатика и кибернетика", доцент кафедры информатики Национального фармацевтического университета (Харьков, Украина).

978-613-9-58671-4

Монография



Марина Нессонова

Математические модели и методы построения классификаторов в медицине

LAP LAMBERT Academic Publishing

Нессонова

Марина Нессонова

Математические модели и методы построения классификаторов в медицине

Марина Нессонова

Математические модели и методы построения классификаторов в медицине

LAP LAMBERT Academic Publishing RU

Imprint
Any brand names and product names mentioned in this book are subject to trademark, brand or patent protection and are trademarks or registered trademarks of their respective holders. The use of brand names, product names, common names, trade names, product descriptions etc. even without a particular marking in this work is in no way to be construed to mean that such names may be regarded as unrestricted in respect of trademark and brand protection legislation and could thus be used by anyone.

Cover image: www.ingimage.com

Publisher:
LAP LAMBERT Academic Publishing
is a trademark of
International Book Market Service Ltd., member of OmniScriptum Publishing Group
17 Meldrum Street, Beau Bassin 71504, Mauritius

Printed at: see last page
ISBN: 978-613-9-58671-4

Zugl. / Утверд.: Киев, Международный научно-образовательный центр информационных технологий и систем, 2015

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

На современном этапе медицинские информационные технологии развиваются достаточно быстрыми темпами. Внедрение единой системы здравоохранения eHealth обусловило повсеместное применение в практике работы медицинских учреждений различных ИТ-решений, что в свою очередь привело к необходимости расширения стандартных функций медицинских информационных систем (МИС). К базовым задачам МИС, таким как ведение электронных медицинских карт пациентов, накопление информации для формирования статистической отчётности, автоматизация документооборота и т.п., в настоящее время добавляются возможности интеллектуального ассистента врача, помогающие выполнять постановку диагноза, проводить анализ динамики патологических процессов и осуществлять прогноз лечения [1—3]. Таким образом, МИС, отвечающие современным требованиям, необходимо включают функции интеллектуальных экспертных систем и систем поддержки принятия решений, в основу разработки которых ложатся математические модели, построенные на базе анализа накопленных данных о состоянии пациентов вследствие различных заболеваний, травматических или иных повреждений органов и систем организма. Современный подход к прогнозированию дальнейшего течения и клинического исхода заболеваний, осуществлению диагностики и оценке тяжести состояния пациентов основан на анализе причинно-следственных связей между характеристиками наблюдаемого состояния и формировании на основе результатов данного анализа решающих правил, согласно

которым осуществляется процесс диагностики [4—7]. Поскольку величины, содержащиеся в массивах медицинских данных, в большинстве своём являются стохастическими, а исследуемые явления и процессы, как правило, протекают в условиях многофакторности, одним из наиболее перспективных подходов к построению математических моделей в медицинских приложениях считается использование методов многомерного статистического анализа [8—10]. Формализация процесса принятия решений врачом возможна посредством исследования закономерностей, которые связывают наблюдаемые характеристики пациента с группой (диагнозом, степенью тяжести, формой или исходом заболевания), к которой он принадлежит, на основе анализа существующих прецедентов. Поэтому при разработке методов и моделей для систем поддержки принятия решений в медицине одним из возможных подходов является решение данной задачи в постановке задачи классификации с обучением (обучения с учителем, или задачи дискриминации).

Разработкой математических моделей, информационных технологий и систем для решения задач классификации с обучением в современных медицинских приложениях занимались такие учёные как Гельфанд И.М., Шифрин М.А., Орлов А.И., Халафян А.А., Леонов В.П., Воронцов К.В. Тем не менее, на протяжении последних 30 лет справедливым считается мнение о том, что математические модели, формализующие и обеспечивающие поддержку принятия решений врачом, должны разрабатываться и корректироваться не только с учётом специфики каждого конкретного заболевания или состояния, а также с течением времени, в зависимости от региона и других факторов [11, 12]. Кроме того, специфика исходных данных в

сложноформализуемых областях человеческой деятельности, к которым относится и медицина, накладывает дополнительные требования и вносит свои особенности к подходам к построению моделей для решения задач данной сферы, которые в первую очередь должны учитывать возможную неполноту информации и позволять совместный анализ множества разнотипных признаков.

Всё вышесказанное обуславливает актуальность и необходимость разработки новых моделей и методов классификации для математического обеспечения интеллектуальных экспертных систем в медицине и врачебных систем поддержки принятия решений.

РАЗДЕЛ 1
СОВРЕМЕННЫЕ ПОДХОДЫ К ПОСТРОЕНИЮ МОДЕЛЕЙ КЛАССИФИКАЦИИ В МЕДИЦИНСКИХ ПРИЛОЖЕНИЯХ

1.1 Постановка задачи классификации с обучением

По сути, с точки зрения прикладной медицинской задачи классификация представляет собой процесс отнесения пациента к одной из заранее заданных групп, определяемых диагнозом либо исходом заболевания, либо формой или степенью тяжести состояния (или т.п.), характеризуемых сходными значениями клинических показателей и симптомов. Формализация данного процесса возможна посредством исследования закономерностей, связывающих наблюдаемые характеристики пациента с группой (классом), к которой он принадлежит, на основании анализа существующих прецедентов (наблюдений). Обычно при разработке методов и моделей классификации исходят из постановки практической задачи обучения по прецедентам (обучения с учителем, или задачи дискриминации) [13]. Приведём одну из возможных формализаций постановки данной задачи (по [13]).

Пусть имеются множество объектов X, множество ответов Y, и существует целевая функция $y^*: \mathrm{X} \to \mathrm{Y}$, значения которой $y_i = y^*(x_i)$ известны только на конечном подмножестве объектов $\{x_1,\ldots,x_L\} \subset X$. Пары «объект–ответ» (x_i, y_i) называются прецедентами. Совокупность пар $\{(x_i, y_i)\}_{i=1}^{L}$ называется обучающей выборкой. Задача обучения по прецедентам заключается в том, чтобы

восстановить функциональную зависимость между объектами и ответами, то есть построить отображение $a: X \to Y$, удовлетворяющее ряду требований, основным из которых является то, что отображение $a(x)$ должно воспроизводить на объектах обучающей выборки заданные ответы: $a(x_i) = y_i$, $\forall i = 1,..L$. Требования выполнения равенства для всех элементов обучающей выборки, как и само равенство, как правило, понимаются как приближённые, хотя в некоторых задачах (например, распознавание образов) бывает необходимо построение так называемых корректных алгоритмов, то есть не допускающих ошибок на обучающей выборке.

Под данную постановку попадают, как задачи восстановления функциональных зависимостей (регрессии), прогнозирования, так и непосредственно задачи классификации. Задачи классификации отличаются тем, что пространство ответов Y принято формально представлять в виде конечного подмножества множества натуральных чисел $Y = \{1,...,m\} \subset N$, где $j = 1,...,m$ – номера классов. В связи с этим более уместно использовать другую постановку задачи [14]:

Пусть имеется множество допустимых объектов X, которое покрыто конечным числом подмножеств (классов) $C_1,....,C_m$: $X = \bigcup_{j=1}^{m} C_j$. Разбиение определено не полностью, а задана лишь некоторая информация I_0 о классах $C_1,....,C_m$ (обучающая информация). Основная задача состоит в том, чтобы по информации I_0 о классах $C_1,....,C_m$ и описанию допустимого объекта $I(x)$ определить (вычислить) значения предикатов $P_j(x) = \{x \in C_j\}$, $j = 1,...,m$.

Основными формами представления обучающей информации являются матрицы (таблицы) типа «объект—признак», или матрицы попарных сравнений объектов.

Матрицы признакового описания объектов (таблицы «объект—свойство» [15], таблицы «объект—признак» [13], таблицы обучения [14]) являются наиболее часто встречающимся видом представления обучающей информации. В этом случае информация I_0 записывается в виде $L \times n$-матрицы:

$$I_0 = \left(f_j(x_i)\right)_{i=1,..L;\, j=1,..n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_L) & \dots & f_n(x_L) \end{pmatrix},$$

где функции $f_j(x)$ являются результатами измерения некоторых характеристик объекта x и называются признаками.

Признаки представляют как функции, действующие из пространства объектов в некоторое пространство значений: $f : X \to D_f$ – область значений признака f. Природа множества значений признака может быть произвольной, но обычно рассматривают признаки следующих типов:

1) Если $D_f = \{a_1,\dots,a_k\}$ – конечное число значений произвольной природы, то говорят о номинальной шкале измерения признака. Частным случаем номинальных являются бинарные (дихотомические) признаки (при $k = 2$), в этом случае область значений принято представлять в виде $D_f = \{0,1\}$ (признак выполнен / не выполнен на объекте).

2) Если $D_f = \{a_1,\dots,a_k\}$ – конечное множество значений с заданным на нём отношением порядка, то признак с такой областью

определения называется порядковым.

3) Если $D_f = (a,b)$, $D_f = (a,b]$, $D_f = [a,b)$ или $D_f = [a,b]$, где $a,b \in R \cup \{\pm\infty\}$, то f — количественный признак.

4) Рассматривают также задачи, в которых D_f представляет собой функции из некоторого класса функций, в частности функции распределения некоторой случайной величины.

В практических приложениях обычно имеют дело с признаками видов 1—3. Если все признаки одного типа, исходная информация называется однородной, в противном случае – разнородной. Наиболее хорошо формализованы процедуры решения задач классификации, в которых обучающая информация является однородной и состоит из количественных признаков. Более того, все подобные методы построения алгоритмов классификации исходят из определённых допущений относительно шкалы измерения, вида функции распределения и других характеристик признаков.

Другой вид представления обучающей информации – таблица попарных сравнений представляет собой квадратную $L \times L$-матрицу $D = \left(d_{ij}\right)_{i,j=1}^{L}$, каждый элемент которой d_{ij} является результатом сопоставления объектов x_i и x_j в смысле некоторого заданного отношения их сходства или различия, предпочтения, взаимосвязи в некотором процессе, которое также может быть выражено в терминах расстояния между объектами, если пространство объектов может быть интерпретировано как метрическое. Подобное представление обучающей информации используется, как правило, при автоматической классификации (без учителя), в методах редукции размерности пространства признаков, и т.п.

В данной работе решается задача классификации с учителем, т.е. обучения по прецедентам, или классификации с обучением, когда пространство ответов Y представляет собой конечной множество областей, на которые разбивается пространство объектов X (классов). При этом обучающая информация I_0 подаётся в виде признакового описания объектов и является разнородной. Таким образом, постановка задачи работы может быть сформулирована в следующем виде:

Есть конечное множество объектов $\{x_i\}_{i=1}^{L} \subset X$, каждый из которых известен по своему описанию с помощью некоторого набора признаков (переменных) $\{f_i\}_{i=1}^{n}$, которые можно представить как функции, действующие из пространства объектов в некоторое пространство значений: $f_i : X \to D_{f_i}$ – область значений признака f_i.

Кроме того известно, что пространство объектов некоторым образом разделено на классы $X = \bigcup \{C_j\}_{j=1}^{m}$ и для каждого объекта из обучающей выборки $\{x_i\}_{i=1}^{L}$ известно, к какому классу он принадлежит.

Требуется построить классификатор (решающее правило, метод, алгоритм классификации, распознаватель), который будет относить объекты к классам на основании значений их признаков $\{f_i\}_{i=1}^{n}$.

1.2 Показатели качества работы классификаторов

Основным показателем качества работы модели классификации является её точность, которая определяется как доля (процент)

правильно распознанных моделью объектов в каждом классе. Общая точность характеризует процент правильно определяемых классификатором случаев во всей обучающей выборке. В случае двухклассовой классификации для оценки качества работы модели часто используются такие показатели как специфичность и чувствительность. Если в качестве примера взять модель классификации, предназначенную для определения наличия некоторого заболевания (т.е. для классификации на два класса: «больные» и «здоровые»), то специфичность можно определить как долю правильно определённых классификатором здоровых людей, а чувствительность – как долю правильно определённых больных. В системах оценки тяжести состояния пациентов под специфичностью понимают процент правильно предсказанных лёгких форм, под чувствительностью – процент правильно предсказанных тяжёлых форм заболевания или состояния [16—20]. В теории классификации с этими понятиями связаны такие определения как ложноположительные, ложноотрицательные, истинно положительные и истинно отрицательные ответы. Рассмотрим эти определения и связанные с ними термины более подробно.

Пусть в обучающей выборке присутствовало P_{real} образцов из первого класса (т.н. «положительные»; например, пациенты с наличием заболевания) и N_{real} образцов из второго класса (т.н. «отрицательные»; пациенты, у которых данное заболевание отсутствует). Пусть также некоторый построенный нами классификатор распознал как положительные только *TP* образцов из первого класса, и как отрицательные только *TN* образцов из второго класса (табл. 1.1). Число *TP* отражает количество

истинно положительных ответов (true positive; число пациентов с заболеванием, которые были правильно распознаны моделью), *TN* – истинно отрицательных (true negative; число реально здоровых людей, которые были правильно отнесены классификатором к классу здоровых).

Число ошибок в первом классе называется ложноотрицательными откликами (*FN*, false negative). Для этого показателя также используются термины «ложные пропуски» и «ошибки второго рода». В нашем примере ошибка второго рода состоит в том, что классификатор не распознаёт (пропускает) реально больного пациента и относит его к классу здоровых.

Таблица 1.1

Характеристики точности модели классификации

		Реальные классы		
		Positive	Negative	
Результат работы классификатора	Positive	*TP*	*FP*	P_{pred}
	Negative	*FN*	*TN*	N_{pred}
		P_{real}	N_{real}	

Число ошибок во втором классе называется ложноположительными откликами (*FP*, false positive), или «ложными обнаружениями», или «ошибками первого рода». Для диагностической системы из нашего примера ошибка первого рода состоит в том, что классификатор ошибочно относит здорового человека к классу больных (обнаруживает заболевание там, где его

реально не существует).

Чувствительность определяется как отношение истинно положительных ответов к общему размеру первого («положительного») класса:

$$Sens = \frac{TP}{P_{real}} = \frac{TP}{TP + FN}.$$

Специфичность – это отношение истинно отрицательных ответов к общему размеру второго («отрицательного») класса:

$$Spec = \frac{TN}{N_{real}} = \frac{TN}{TN + FP}.$$

Уровень ошибки I рода вычисляется как доля (или процент) ложноположительных образцов во втором («отрицательном») классе:

$$\alpha = \frac{FP}{N_{real}} = \frac{FP}{TN + FP} = 1 - Spec.$$

Уровень ошибки II рода вычисляется как доля (или процент) ложноотрицательных образцов в первом («положительном») классе:

$$\beta = \frac{FN}{P_{real}} = \frac{FN}{TP + FN} = 1 - Sens.$$

Таким образом, под определениями «высокочувствительная диагностическая система», «высокочувствительный тест» и т.п. в медицинских приложениях понимается модель классификации, способная на основании используемых в ней показателей определять наличие заболевания у реально больных людей с высокой точностью. Под «высокоспецифичной системой» подразумевается классификатор, способный с высокой точностью распознавать отсутствие заболевания у людей, которые реально здоровы. Синонимом высокой чувствительности является наличие низкой ошибки II рода, а специфичности – наличие низкой ошибки I рода.

Очевидно, хороший классификатор должен обладать и высокой специфичностью, и высокой чувствительностью, т.е. одинаково хорошо распознавать оба класса, однако на практике совместить эти два требования удаётся не всегда. В этом случае в практических приложениях при построении моделей классификации большинство источников советуют исходить из специфики решаемой задачи. Например, в некоторых случаях ошибочное диагностирование наличия заболевания сопряжено с высоким риском побочных эффектов и осложнений для пациента, либо дорогостоящей терапией. В таких случаях высокоспецифичную модель классификации лучше предпочесть высокочувствительной. Если же гипердиагностика не является критичной (например, при заболеваниях и состояниях, где окончательный диагноз должен быть подтверждён дополнительными исследованиями), то предпочтительными являются высокочувствительные классификаторы.

Способом соблюсти баланс между специфичностью и чувствительностью является оценка качества работы классификаторов с помощью анализа ROC-кривых. Кривая ROC (receiver operating characteristic curve) показывает зависимость чувствительности от уровня ошибки I рода при изменении порогового значения в решающем правиле классификации и является идеальным наглядным способом сравнительной оценки качества работы нескольких моделей (рис. 1.1). Так, на приеденном рисунке ROC-кривая для модели 1 (показанная сплошной линией) расположена выше ROC-кривой модели 2 (показанной в виде точек), что отражает тот факт, что модель 1 имеет более высокий уровень верных обнаружений (чувствительность) при более низком уровне ошибок I рода. Таким образом, модель 1 явно предпочтительнее модели 2 по качеству

классификации.

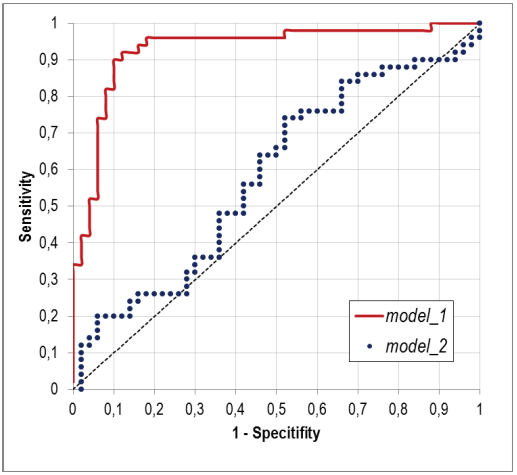


Рис 1.1. ROC-кривые для двух моделей классификации

Опорная линия на рис. 1.1 (биссектриса координатного угла, показанная пунктиром) соответствует гипотетическому случаю, когда чувствительность классификатора равна уровню ошибки I рода. Очевидно, что такой классификатор, также как и классификатор, ROC-кривая которого лежит ниже опорной линии, является бесполезным.

Наиболее часто используемой количественной мерой качества классификатора в ROC-анализе является AUC (area under curve) – площадь под его ROC-кривой, вычисляемая как интеграл от чувствительности по уровню ошибки I рода:

$$AUC = \int_0^1 Sens\, d\alpha .$$

Для классификатора, 100%-но точно разделяющего классы,

$AUC = 1$, для классификатора, имеющего ROC-кривую, совпадающую с опорной линией, $AUC = 0{,}5$. Шкала для качественной характеристики точности классификатора, основанная на значении AUC, предлагаемая во многих источниках (например, [21, 22]), приведена в первых двух колонках табл. 1.2. Альтернативной мерой служит коэффициент Джини (Gini), являющийся удвоенной площадью между ROC-кривой и опорной линией. В некоторых случаях коэффициент Джини более удобен тем, что диапазон его изменения от 0 до 1, в то время как значения AUC рассматриваются в интервале от 0,5 до 1. Коэффициент Джини может быть вычислен через площадь под ROC-кривой по формуле:

$$Gini = \left(AUC - \frac{1}{2} \right) \times 2 = 2 \cdot AUC - 1 .$$

Шкала для качественной характеристики точности классификатора, основанная на значении коэффициента Джини, приведена в последних двух колонках табл. 1.2. Если в качестве примера взять два классификатора, ROC-кривые которых показаны на рис. 1.1, то для модели 1 (сплошная линия) площадь под кривой $AUC = 0{,}930$, а коэффициент Джини $Gini = 0{,}860$, что говорит об отличном качестве классификации, предоставляемом данной моделью. В то же время для модели 2 (точечная линия) $AUC = 0{,}583$, а коэффициент Джини $Gini = 0{,}166$, что свидетельствует о плохом качестве этого классификатора. Графики ROC-кривых наглядно отражают этот факт.

Более детально вопросы применения анализа ROC-кривых в медицинских приложениях рассмотрены в работах [19, 23, 20, 22, 21].

Обратим внимание, что рассмотренные выше характеристики качества работы моделей классификации оперируют показателями,

оцениваемыми при известных реальных классах. Т.е. специфичность, чувствительность и уровни ошибок I и II рода вычисляются как отношения правильно/ошибочно распознанных объектов к количеству объектов в том или ином классе, которое должно быть задано заранее. Таким образом, эти показатели характеризуют качество работы классификатора на известных данных, другими словами, его *точность покрытия* классов. В то же время при применении разработанной модели классификации на новых данных количество объектов в каждом классе заранее неизвестно, поэтому для оценки качества её будущей работы, т.н. *прогностической точности*, более уместным кажется использование других показателей.

Таблица 1.2

Оценка качества модели классификации по значениям площади под ROC-кривой (AUC) и коэффициента Джини (Gini)

Диапазон значений AUC	Качество классификации	Диапазон значений коэффициента Gini
0,9 ÷ 1,0	Отличное	0,8 ÷ 1,0
0,8 ÷ 0,9	Хорошее	0,6 ÷ 0,8
0,7 ÷ 0,8	Среднее	0,4 ÷ 0,6
0,6 ÷ 0,7	Слабое	0,2 ÷ 0,4
0,5 ÷ 0,6	Неудовлетворительное	0 ÷ 0,2

Так, аналогично понятию чувствительности вводится понятие точности положительных предсказаний (positive predictive value, *PPV*, или precision) [24—27], отражающей долю правильно распознанных образцов из первого («положительного») класса среди всех объектов,

отнесённых классификатором к этому классу:

$$PPV = \frac{TP}{TP + FP}.$$

Очевидно, «хороший» сбалансированный классификатор должен иметь как высокую точность покрытия (чувствительность), так и высокую прогностическую точность (*PPV*). Эти два показателя равновзвешено используются в F_1-мере качества классификации:

$$F_1 = \frac{2}{\frac{1}{TPV} + \frac{1}{Sens}} = \frac{2 \cdot TPV \cdot Sens}{TPV + Sens}.$$

Если же точности покрытия и прогностической точности при решении конкретной прикладной задачи придаётся разная важность, то для оценки качества классификатора предлагают использовать обобщённый показатель F_ω-меры, в котором чувствительность имеет вес ω, а точность положительных предсказаний вес $(1-\omega)$:

$$F_\omega = \frac{2}{\frac{1-\omega}{TPV} + \frac{\omega}{Sens}} = \frac{(1-\omega) \cdot TPV \cdot Sens}{\omega \cdot (1-\omega) \cdot TPV + Sens}.$$

Аналогично вводится показатель *NPV* (negative predictive value) для оценки прогностической точности во втором («отрицательном») классе, являющийся отношением количества правильно распознанных «отрицательных» образцов к общему количеству объектов, спрогнозированных моделью классификатора как «отрицательные», а также мера баланса между точностью отрицательных предсказаний и специфичностью.

В [28] для различения понятий точности покрытия и прогностической точности были введены, на наш взгляд, достаточно удачные определения. Так, специфичность и чувствительность, оцениваемые по известным данным, были названы термином

точность производителя (producer's accuracy), в то время как точности положительных и отрицательных предсказаний (*PPV* и *NPV*) – *точностью пользователя* (user's accuracy).

Таким образом, в данном параграфе рассмотрены основные подходы к оценке качества работы моделей классификации и используемые для этого показатели. Изложенный материал не претендует на исчерпывающую полноту и всесторонность освещения вопроса, в практических приложениях могут использоваться и другие меры для оценки качества классификации, являющиеся обобщением, развитием или уточнением рассмотренных индексов.

1.3 Методы решения задачи классификации с учителем

1.3.1 Методы построения моделей классификации с учителем при использовании определённых ограничений свойств исходных данных

Для решения задачи классификации с учителем существует ряд методов, некоторые из которых исторически возникли достаточно давно и на сегодняшний день считаются классическими и общепринятыми. К таким методам можно отнести дискриминантный анализ (ДА), основные положения которого были выведены и сформулированы Р. Фишером в [29]. Понятие ДА настолько тесно ассоциировалось с задачей классификации с обучением, что во многих работах саму задачу называют задачей дискриминантного анализа, или задачей дискриминации (см., например, [30]). Однако под ДА понимается вполне конкретный метод построения решающих правил, называемых дискриминантными функциями, позволяющих относить объекты к двум или более классам. В практических приложениях

чаще всего используется классическая модель ДА, в которой дискриминантные функции являются линейными комбинациями объясняющих числовых переменных.

Для оценки степени тяжести состояния пациентов математические модели ДА используются довольно часто. Например, предложена математическая модель прогноза исхода закрытой травмы печени, использующая данные клинико-лабораторных, рентгенологических, ультразвуковых и инструментальных методов исследования [31]. Авторы заявляют о высокой чувствительности (95,83%) предсказания вероятности летального исхода у 110 пострадавших на основании классификационных функций, полученных с помощью линейного дискриминантного анализа Фишера. Однако авторами при построении модели не учтён целый ряд важных требований к исходным данным, при которых линейный ДА Фишера может быть успешно применён. В данной модели использованы только дихотомические объясняющие переменные, тогда как применение метода ДА требует, чтобы входные данные были измерены в интервальной (или, как минимум, в ординальной) шкале [15, 32]. Соответственно, невозможно выполнение и других требований, делающих адекватной модель линейного ДА. К этим требованиям относятся многомерный нормальный закон распределения показателей в классах и равенство ковариационных матриц. Это ставит под сомнение заявленные авторами показатели точности предлагаемой модели и делает её неприменимой при использовании новых данных.

Предложена модель, разработанная методом линейного дискриминантного анализа, для оценки степени тяжести начальных стадий хронической сердечной недостаточности у подростков [33].

Модель использовалась для классификации подростков на три группы: 1 – подростки с заболеваниями сердечнососудистой системы, но без признаков хронической сердечной недостаточности; 2 – подростки с заболеваниями сердца и начальной стадией хронической сердечной недостаточности; 3 – практически здоровые подростки. В статье приведены значения коэффициентов дискриминантных функций и проведён их детальный анализ. Сообщается, что общая точность разработанной модели на обучающих данных 98,1%, однако отсутствуют сведения о точности прогнозирования каждой из степеней тяжести состояния. Также не приведены формулы для вычисления функций классификации, что существенно усложняет применение модели в медицинской практике и её проверку на новых данных.

В работе [34] линейный дискриминантный анализ использован для построения модели прогнозирования периоперационных осложнений в реконструктивной хирургии пищевода. Авторами приводятся две классификационные функции для низкой вероятности развития осложнений (K_1) и высокой вероятности развития осложнений (K_2):

$$\begin{aligned} K_1 = {} & 7{,}71 \times (\textit{вид питания}) + 0{,}95 \times (\textit{точка доступа питания}) - \\ & - 5{,}6 \times (\textit{алкоголизм}) + 0{,}19 \times (\textit{потеря массы тела (кг)}) + \\ & + 0{,}06 \times (\textit{дефицит массы тела (\%)}) + \\ & + 1{,}72 \times (\textit{ИБС, ПИКС, нарушения ритма}) - 8{,}23 \end{aligned},$$

$$\begin{aligned} K_2 = {} & 3{,}76 \times (\textit{вид питания}) + 2{,}90 \times (\textit{точка доступа питания}) - \\ & - 0{,}11 \times (\textit{алкоголизм}) + 0{,}34 \times (\textit{потеря массы тела (кг)}) - \\ & - 0{,}04 \times (\textit{дефицит массы тела (\%)}) - \\ & - 0{,}12 \times (\textit{ИБС, ПИКС, нарушения ритма}) - 7{,}63 \end{aligned}.$$

Оценка точности модели, проведенная на независимой выборке, показала 83,2% правильных предсказаний, однако этот процент точности достигается за счёт высокой специфичности модели (точность предсказания отсутствия осложнений составляет 96%), в то время как чувствительность очень низкая (правильно было определено только 46,2% случаев осложнений).

К недостаткам модели можно отнести и то, что в классификационные функции включены номинальные переменные (вид питания, наличие ИБС, алкоголизма и др.), что является некорректным при использовании линейного ДА.

Вторым, достаточно популярным методом классификации является логистическая регрессия, применяемая в случае двухклассовой классификации. Методы логит- и пробит-регрессии моделируют бинарную выходную переменную (метки двух классов задаются как 0 и 1) как непрерывную переменную со значениями на интервале [0; 1]. Логистическая регрессия позволяет построить линейную модель зависимости от входных переменных величины $y = \ln \frac{p}{1-p}$, являющейся логит-преобразованием вероятности p того, что объект принадлежит классу 1. Таким образом, вероятность находится как $p = \left(1 + e^{-y}\right)^{-1}$. При $p < 0{,}5$ прогнозируют, что объект принадлежит классу 0, при $p \geq 0{,}5$ – классу 1.

Эта модель была использована для прогнозирования риска развития микрососудистых осложнений при сахарном диабете 1-го типа [35]. Авторы построили несколько моделей логистической регрессии для прогнозирования появления поздних осложнений, таких как:

1) Диабетическая ретинопатия (ДР). Точность определения больных с этим осложнением – 81%, без него – 77,4%; общая – 79,5%.
2) Диабетическая нефропатия (ДН). Точность определения больных с этим осложнением – 58,6%, без него – 79,5%; общая – 69,6%.
3) Диабетическая периферическая полинейропатия (ДПН). Точность определения больных с этим осложнением – 75,8%, без него – 80,7%; общая – 78,1%.
4) Диабетическая автономная нейропатия (ДАН). Точность определения больных с этим осложнением – 33,8%, без него – 86,9%; общая – 68,6%.

Очевидно, что удовлетворительная общая точность и чувствительность достигается при прогнозировании только двух типов осложнений (ДР и ДПН) из четырёх.

В работе [36] построена модель логистической регрессии для прогнозирования исхода острого панкреатита, использующая в качестве независимых переменных агрегационные показатели тромбоцитов: время достижения максимальной степени агрегации (x_1) и максимальная скорость агрегации (x_2) с тромбоцитарно-коагуляционными показателями и факторами, связанными с обменными и интоксикационными процессами. В соответствии с данной моделью вероятность летального исхода (X) определяется по формуле:

$$X = \exp(53{,}3 - 2{,}4 \cdot x_1 + 1{,}2 \cdot x_2) / [1 + \exp(53{,}3 - 2{,}4 \cdot x_1 + 1{,}2 \cdot x_2)].$$

Вместе с тем авторами указывается, что высокая точность прогнозирования летального исхода (до 93%) достигается только при одновременном сочетании снижения максимальной степени агрегации тромбоцитов более чем на 30% и замедлении максимальной скорости

агрегации более чем на 100% / мин с третьих суток заболевания. В остальных случаях точность модели – около 70%.

В работе [37] метод логистической регрессии использован для построения шкалы прогнозирования инфицированного панкреонекроза. Авторы заявляют о 90% точности прогноза инфицирования очагов панкреонекроза с помощью построенной модели логит-регрессии, включающей 6 показателей. Однако на практике предлагают использовать не уравнение регрессии, в котором эти показатели учитываются с различными коэффициентами, а присваивать 0 баллов значениям, характерным для стерильного панкреонекроза, а значениям, характерным для инфицированного панкреонекроза, – 1 балл. После подсчёта суммы баллов при значении 0—1 балл предлагается прогнозировать стерильный панкреонекроз; 2 балла – прогноз сомнительный; 3 и более баллов соответствует инфицированию очагов панкреонекроза на протяжении госпитализации.

При использовании моделей логит- и пробит-регрессии для классификации на более двух классов возможно построение нескольких моделей регрессии, осуществляющих последовательную двухклассовую классификацию, как это было сделано, например, в [38], где метод логистической регрессии был использован для определения состава трёх кластеров, включающих экзогенные факторы и сходные группы пациентов с нарушениями мозгового кровообращения.

1.3.2 Методы построения моделей классификации с учителем, свободные от требований к свойствам исходных данных

Упомянутые выше методы ДА и логистической регрессии для решения задачи классификации с обучением являются наиболее широко используемыми, что в бóльшей степени связано с более ранним их появлением. Их разработка и развитие происходили в рамках классического направления прикладной математической статистики в соответствии с «англосаксонской» традицией анализа данных [39, 40]. В этой парадигме математическая модель изучаемого процесса или явления считается известной с точностью до параметров. Т.е. предполагается, что известен набор предикторов (объясняющих переменных), влияющих на выходную (целевую) характеристику, а также общий вид аналитического выражения (уравнения или системы уравнений или неравенств), описывающего взаимосвязь между объясняющими и целевой переменными. Проблема «идентификации модели» состоит в реализации вычислительных процедур, позволяющих определить неизвестные параметры (коэффициенты уравнения) по данным обучающей выборки, содержащим значения целевой переменной при различных значениях объясняющих переменных [41]. Вычисление значений параметров модели осуществляется на основе принципа минимизации ошибки, достигаемой на эмпирических данных. При этом выдвигается ряд гипотез относительно свойств исходных данных, для проверки которых используются данные обучающей выборки. При условии истинности этих гипотез становится возможным идентифицировать такую модель, которая позволяет с заданной точностью определять значения одной (целевой) переменной по значениям других (входных)

характеристик.

Следует отметить, что при построении математических моделей в сложноформализуемых областях человеческой деятельности, к которым относится и медицина, классические инструменты математической статистики нередко оказываются малоэффективными. Это связано со спецификой исходных данных, в которых часто присутствуют пропущенные значения, а из-за отсутствия стандартов кодирования медицинской информации одни и те же показатели могут быть записаны в разных шкалах, например, когда значения интервальных переменных переводятся в более бедную шкалу [8]. Кроме того, требуемые для построения подобных моделей от исходных данных свойства (например, нормальность распределения, определённые свойства корреляционных и ковариационных матриц, однородность дисперсий, и т.п.) в случае таких «некачественных» данных выполняются далеко не всегда. В этой связи актуальным становится подход, лежащий в основе «французской» школы анализа данных, основная идея которого выражена Ж.-П. Бензекри [42] в формулировке: «Модель должна соответствовать данным, а не наоборот». При этом подходе не требуется никаких априорных предположений о свойствах исходных данных, а идентификация модели и обнаружение закономерностей происходит через исследование и анализ их структуры.

Последняя идеология, в частности, легла в основу такого популярного на сегодняшний день направления анализа данных как KDD (Knowledge Discovery in Databases – обнаружение знаний в базах данных) и его основного элемента – технологии Data Mining (добычи данных). В математической статистике под методами Data Mining понимают методы анализа данных при отсутствии предварительно

чётко сформулированных гипотез об их структуре. В [43] Data Mining определяется как процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Часто этот термин используется в качестве синонима разведочного анализа данных [44]. Однако, более строго, разведочный анализ можно представить как совокупность элементов OLAP (online analytical processing – оперативной аналитической обработки данных), т.н. «грубого» разведочного анализа, и непосредственно методов Data Mining («раскопки» данных).

Методы, используемые технологией Data Mining, основываются на концепции поиска шаблонов (паттернов), представляющих собой закономерности, свойственные подвыборкам данных, которые отражают фрагменты многоаспектных взаимоотношений, присутствующие в них, и могут быть компактно выражены в понятной человеку форме [45, 46].

Основные положения, из которых исходят принципы Data Mining, это интерпретируемость, конкретность и понятность результатов, а также нетривиальность получаемых шаблонов, которые должны отражать неочевидные закономерности, скрытые в сырых данных, учитывая при этом разнородность исследуемых признаков.

Разнообразие методов и алгоритмов, используемых Data Mining, обусловлено мультидисциплинарностью этой области, которая возникла и развивается на базе достижений прикладной статистики, распознавания образов, методов искусственного интеллекта, теории баз данных, экспертных систем и т.д. Наиболее удачные методы, как правило, основаны на комбинировании и интеграции сразу

нескольких подходов.

Одним из наиболее популярных подходов к решению задач Data Mining являются деревья решений (или деревья классификации, decision (classification) trees). Они представляют собой структуры логических правил классификации типа «Если... То...», позволяющие интерпретировать шаблоны данных с целью их распознавания. Деревья решений организованы в виде иерархической структуры, состоящей из узлов принятия решений по оценке значений определённых переменных для прогнозирования результирующего значения. Применение деревьев классификации приводит к получению символического обозначения класса в результате последовательных ответов «Да» или «Нет» на ряд вопросов. По существу, при каждой проверке условия происходит сортировка подвыборок данных таким образом, что каждый элемент данных определяется как соответствующий только одному разветвлению. Таким образом, критерии принятия решений разбивают одно общее множество данных на набор непересекающихся подмножеств. В результате объединения таких проверок в древовидную иерархию фактически организуется процесс разбиения всех данных на всё меньшие части, происходящий до тех пор, пока не достигается конечный (листовой) узел. Каждый листовой узел соответствует небольшой, но исключительной (неповторяющейся) части исходного множества [46].

Во многих медицинских приложениях метод деревьев решений показывает хорошие результаты. Например, в работе [47] сообщается о построении дерева решений для оценки степени тяжести (стадии) глаукомы. Построенная модель для определения одной из четырёх стадий заболевания (отсутствие, начальная, развитая и далеко

зашедшая стадии) показала общую точность 97%, специфичность 94,3% и чувствительность 97%.

Дерево решений, построенное авторами в [48], позволяет предсказывать исход эндовенозной лазерной облитерации с точностью 81,1%.

Успешно был применён метод деревьев решений для классификации пациентов по степени тяжести гипертонической болезни при сахарном диабете 2-го типа [49]. Общая точность разработанной модели – 92%, специфичность – 93,1%, чувствительность – 90,5%.

Хорошие результаты получены методом деревьев классификации при оценке эффективности восстановительного лечения [50], а также при диагностике хронической сердечной недостаточности пациентов с ишемической болезнью сердца [51].

В работе [52] автором разработана модель дерева классификации для прогнозирования развития осложнений деструктивного панкреатита (жидкостных полостных образований поджелудочной железы), обладающая точностью 96,5%.

При разработке системы интеллектуального анализа данных для прогнозирования результатов хирургического лечения атеросклероза [53] авторами проведён сравнительный анализ методов решающих деревьев, логистической регрессии, опорных векторов и нейронных сетей и показано преимущество деревьев решений перед другими использованными методами.

Деревья классификации идеально приспособлены для графического представления и поэтому, сделанные на их основе выводы гораздо легче интерпретировать, чем, если бы они были представлены только в числовой или описательной текстовой форме.

Лёгкостью интерпретации результатов, наглядностью, интуитивностью и понятностью в большей степени объясняется популярность и широкая применимость данного метода. К недостаткам метода критики относят то, что деревья решений принципиально не способны находить «лучшие», т.е. наиболее полные и точные, паттерны в данных, т.к. они реализуют наивный принцип последовательного просмотра признаков, что иногда приводит к выявлению только «осколков» настоящих закономерностей [54].

Другой класс методов, активно используемых Data Mining, – это системы рассуждений на основе прецедентов (case based reasoning, CBR). Рассуждения на основе прецедентов основываются на накоплении опыта и последующей адаптации решения известной задачи к решению новой. Прецедентный подход позволяет упростить процесс принятия решений в условиях временных ограничений и при наличии различного рода неопределённости в исходных данных и экспертных знаниях [55]. Для того чтобы сделать прогноз на будущее или выбрать правильное решение, эти системы на основании всей накопленной информации находят в памяти аналоги, близкие к имеющейся ситуации, и выбирают тот же ответ, который был для них правильным. Данный метод известен также под названием метода «ближайшего соседа» (nearest neighbour). Понятие близости образцов в разных задачах может трактоваться по-разному. Для оценки близости существует множество мер, и от правильного выбора меры существенным образом зависит качество классификации или прогноза, а также объём множества прецедентов, которые нужно хранить в памяти [56]. Системы CBR показывают неплохие результаты в самых разнообразных задачах, однако главным их

минусом считают то, что они вообще не создают каких-либо моделей или правил, обобщающих предыдущий опыт, — в выборе решения они основываются на всём массиве доступных исторических данных, поэтому невозможно сказать, на основе каких конкретно факторов был построен ответ [54].

Хотя практически во всех обзорах по интеллектуальному анализу данных медицинская диагностика упоминается как одно из перспективных приложений систем CBR, при анализе доступных публикаций нами выявлено очень мало сведений о каких-либо конкретных примерах их применения для решения практических задач. Примером могут служить работы [57, 58], в которых сообщается о создании и результатах использования программно-информационной системы диагностики состояния в случае ревматических заболеваний.

Довольно большой класс систем, используемых в задачах классификации, распознавания и восстановления зависимостей, составляют искусственные нейронные сети. Они представляют собой системы, состоящие из большого числа искусственных нейронов, соединённых между собой и с внешней средой с помощью связей, определяемых весовыми коэффициентами [59]. Схема искусственного нейрона показана на рис. 1.2 [60].

Входные значения (сигналы) усиливаются или ослабляются в зависимости от величины весовых коэффициентов, затем их взвешенная сумма служит аргументом для некоторой функции активации. Искусственные нейроны различаются в зависимости от используемых функций активации, которые, как правило, выбираются из семейства пороговых, (кусочно-) линейных или сигмоидальных, функций. Нейронные сети различаются способом объединения

искусственных нейронов в слои, т.е. архитектурой.

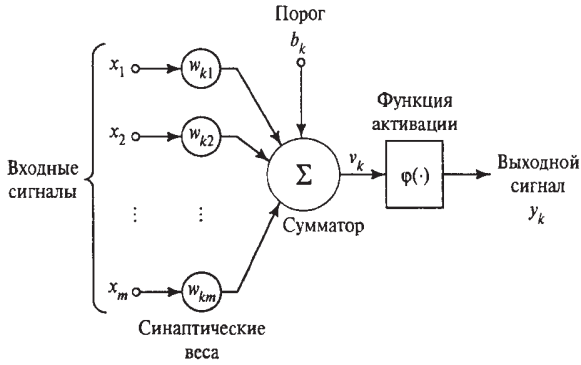


Рис. 1.2. Схема искусственного нейрона [60]

Выбор определённой архитектуры искусственной нейронной сети обуславливает алгоритмы, которые будут использоваться для её обучения. Процесс обучения состоит в определении таких значений весовых коэффициентов межнейронных связей, которые обеспечивали бы наилучшее качество её работы, для чего используется обучающая выборка наблюдений, для которых известны значения, как входных параметров, так и соответствующие им правильные ответы.

К основным недостаткам моделей нейронных сетей относят необходимость наличия очень большого объёма обучающей выборки, а также то, что даже натренированная нейронная сеть представляет собой т.н. «чёрный ящик». Знания, зафиксированные как веса нескольких сотен межнейронных связей, совершенно не поддаются анализу и интерпретации человеком, а известные попытки дать интерпретацию структуре настроенной нейросети выглядят

неубедительными [54, 50, 46]. Одна из наиболее часто применяемых в практических приложениях модель нейросетевой архитектуры (многослойный персептрон) обладает высокой вычислительной мощностью. В то же время наличие скрытых слоёв, на которых процесс обучения трудно отследить, распределённая форма нелинейности и высокая связность сети служат причиной недостаточной исследованности (и теоретической, и практической) подобных моделей [60]. Кроме того, многослойный персептрон довольно неустойчив к помехам в исходных данных, что является характерным и для других моделей нейронных сетей [60, 61].

Тем не менее, применение нейросетевых моделей в медицине на сегодняшний день достаточно популярно, что, возможно, объясняется понятными для специалистов в медицине аналогиями с работой нейронов мозга, нервной ткани и т.п., с которых обычно начинается объяснение идеологической парадигмы возникновения метода. Так, в работе [62] на основе нейронных сетей разрабатывается система прогнозной диагностики транзиторных ишемических атак (общая точность – 78%, специфичность – 89%, чувствительность – 73%).

В [63] сообщается о построении и результатах использования нейросетевой модели для прогнозирования исходов хирургического вмешательства и течения послеоперационного периода у больных, перенесших операцию протезирования клапанов сердца, чувствительность и специфичность которой составили 74 и 62% соответственно.

Преимущество модели нейронной сети, построенной для прогноза нарушений психической адаптации, перед регрессионной моделью и моделью дискриминантного анализа показано в работе [64].

В [65] авторы приводят результаты построения нейросетевой модели прогнозирования вероятности развития инфицированного панкреонекроза, точность которой составляет 90%, а специфичность – 96%. В работе [66] тех же авторов приведен обзор иностранных источников, свидетельствующих о применении методов нейросетевого моделирования в различных областях медицинской науки.

Для оценки тяжести послеоперационного состояния при радикальных операциях по поводу рака лёгкого в [67] проведено построение и исследование ряда моделей логистической регрессии и нейронных сетей. Как демонстрируют приведенные результаты, во многих случаях модели логит-регрессии показывали лучшее качество работы на тестовых выборках, чем нейронные сети.

Не последнее место в технологиях Data Mining принадлежит методам визуализации многомерных данных. Если на этапе грубого разведочного анализа эту визуализацию возможно осуществить с помощью двух- или трёхмерных графиков рассеивания, диаграмм распределения и размаха, а также их матриц, то в случае большого количества исследуемых переменных информативность восприятия и возможность анализа таких представлений существенно снижается. Для выхода из ситуации применяются всевозможные методы преобразования пространства данных с целью снижения его размерности при максимально возможном сохранении связей, существующих в исходном массиве данных. Достаточно детально подходы к решению этой проблемы изложены в монографии [68].

Довольно мощными методами визуализации многомерных данных с целью выявления скрытых в них ассоциаций и закономерностей являются методы многомерного шкалирования и

анализа соответствий (корреспондентского анализа). К сожалению, применение этих методов в медицинских приложениях можно встретить достаточно редко, а результаты их применения не подвергаются содержательному анализу.

Например, в работе [69] анализ соответствий применён для исследования комбинированного применения 10 групп антибиотиков при лечении пневмонии. В [70] с помощью анализа соответствий изучалась взаимосвязь между диспластическим синдромом или фенотипом и эзофагогастродуоденальными заболеваниями. В работе [71] рассматриваются перспективы применения многомерного шкалирования для электроэнцефалографического мониторинга психофармакотерапии, а в [72] теми же авторами сообщается о результатах его применения. Простой корреспондентский анализ применён в [73] для исследования связей между факторами, влияющими на смертность пациентов отделений интенсивной терапии, а в [74] – для получения профилей кардиологических пациентов с диабетом, наиболее подверженных стрессам и депрессиям.

Таким образом, анализ существующих методов решения задачи классификации с учителем показывает, что наиболее перспективными в современных прикладных задачах анализа медицинских данных являются методы, свободные от требований к свойствам исходных данных. При этом построение математической модели какой-либо конкретной прикладной задачи должно исходить из анализа и исследования структуры накопленных по этой тематике данных, т.е. детального анализа структуры и свойств обучающей информации. Причём методы этого анализа не должны быть ограничены какими-

либо априорными предположениями о виде модели и жёсткими рамками, необходимыми для её реализации. Это обусловило выбор направления исследований, результаты которых изложены в данной монографии.

На основании проведенных в данном разделе исследований можно сделать следующие выводы:

1. Анализ последних публикаций показал, что при использовании в медицинских приложениях классических методов статистического моделирования на практике часто игнорируются требования к исходным данным, лежащие в основе разработки этих методов. Подобное игнорирование приводит к созданию неадекватных моделей, применение которых к новым данным не даёт ожидаемых результатов.

2. Анализ существующих методов построения моделей классификации с обучением позволил заключить, что наиболее перспективными в медицинских приложениях являются методы, свободные от требований к исходным данным, что обусловлено спецификой исходных данных в подобных практических задачах.

3. В результате исследования источников литературы и публикаций, касающихся использования методов KDD для исследования данных и построения моделей в прикладных задачах медицины, обнаружен целый класс методов – методы графической интерпретации и визуализации многомерных данных, – перспективы и результаты применения которых ещё недостаточно изучены.

РАЗДЕЛ 2
МЕТОДИЧЕСКИЕ ОСНОВЫ РАЗРАБОТКИ ИНФОРМАЦИОННОЙ ТЕХНОЛОГИИ ДЛЯ РЕШЕНИЯ ЗАДАЧ КЛАССИФИКАЦИИ В МЕДИЦИНСКИХ ПРИЛОЖЕНИЯХ

2.1 Определение этапов разработки информационной технологии, их задач и методов решения

Целью работы является разработка информационной технологии реализации моделей классификации для медицинских приложений, базирующейся на методах классификации с обучением. Выбор, использованных в работе подходов и методов, был обусловлен следующими этапами, необходимыми достижения поставленной цели. Так, для построения классификатора по обучающей информации требовалось:

1. На этапе 1 – оценить степень влияния каждого из признаков, описывающих объекты, на принадлежность объектов к различным классам.

2. На этапе 2 на основании результатов проведенной на этапе 1 оценки разработать правила отнесения объектов к классам.

3. На этапе 3 при необходимости сформировать комбинацию из нескольких классификаторов для достижения лучшего качества распознавания классов.

Схематично этапы разработки информационной технологии и методы, использованные для решения задач работы, представлены на рис. 2.1.

Рис. 2.1. Этапы разработки информационной технологии, их задачи и методы решения

Первый этап разработки метода решения задачи классификации с обучением должен состоять в анализе структуры обучающей информации с целью извлечения из неё сведений о закономерностях и ассоциациях, характерных для специфики решаемой задачи. Учитывая разнородность и многомерность входных данных прикладных задач медицины, а также необходимость удобства анализа и репрезентации извлечённой информации, наиболее целесообразным на данном этапе представляется применение методов геометрической интерпретации многомерных данных. Эти методы в настоящее время используются на этапе разведочного анализа исключительно в качестве инструмента описания данных, позволяющего посредством визуализации обнаружить взаимосвязи отдельных показателей и/или их категорий значений или выделить сходные подгруппы в данных. Обычно применение какого-либо метода графической интерпретации структуры данных завершается словесным описанием полученного визуального представления. Однако, на наш взгляд, потенциал

применения этой группы методов многомерного анализа недостаточно использован. Одной из целей в данной работе было развить эти методы и использовать полученную графическую интерпретацию для создания формальных правил, описывающих структуру данных, а именно, правил, формализующих отнесение объектов к нескольким наперёд заданным подгруппам (классам) в имеющейся обучающей информации.

Итогом применения какого-либо метода геометрической интерпретации структуры данных является карта, графически представляющая признаковое описание объектов. По этой карте можно визуально оценить силу взаимосвязей между объектами, признаками или категориями их значений, трактуя её как степень близости точек, их представляющих.

Одной из проблем, которая может возникнуть на этом этапе, является то, что при нанесении на карту достаточно большого количества объектов возможность её содержательного, удобного и адекватного анализа существенно снижается. Кроме того, наличие в модели классификатора переменных, не оказывающих влияния на принадлежность объектов к классам, или дублирующих друг друга (т.е. несущих одинаковую информацию или оказывающих равноценный вклад), противоречит принципу регуляризации [13, 75]. При небольшой размерности матрицы обучения этой проблемы может и не возникнуть. Однако в реальных задачах количество анализируемых признаков, как правило, достаточно велико, и поэтому возникает проблема их сокращения, выбора из общего набора только тех, которые будут действительно полезными для объяснения правил формирования классов. Поэтому в качестве вспомогательного шага,

предваряющего первый этап, необходимо решить задачу формирования неизбыточного набора показателей, значимо влияющих на принадлежность объектов к классам. В данной работе эта задача была решена путём включения в модель только тех признаков, для которых на обучающей выборке подтверждалось их статистически значимое различие между классами. С целью обнаружения этого различия применялись методы парных и множественных сравнений для количественных признаков и анализ таблиц сопряжённости качественных признаков.

Для получения геометрической интерпретации взаимосвязей между классами объектов и объясняющими признаками рассмотрены перспективы использования методов многомерного шкалирования и множественного корреспондентского анализа.

Для формирования решающих правил отнесения объектов к классам (этап 2) использован метод построения алгоритмов вычисления оценок.

При составлении композиций классификаторов (этап 3) использовались подходы, лежащие в основе принципа специализации и взвешенного голосования.

В последующих параграфах детально рассмотрены методы, использованные на каждом из этапов разработки информационной технологии.

2.2 Методы геометрической интерпретации структуры данных

Методы геометрической интерпретации структуры данных использованы в работе на основном этапе разработки метода решения

задачи классификации с обучением. Эти методы применяются для представления множества объектов в виде точек в пространстве сниженной размерности таким образом, чтобы похожие объекты представлялись близко расположенными точками [76]. Общее назначение этой группы методов – сокращение размерности пространства исследуемых признаков и наглядное отражение взаимоотношений между наблюдениями и переменными [77]. Методы графической интерпретации структуры данных разрабатывались и применяются на практике для исследования сложных явлений и процессов, не поддающихся непосредственному описанию или моделированию, их целью является выявление латентных факторов, определяющих различия или сходство объектов [78].

Наиболее простым и часто используемым методом графической интерпретации является график рассеивания, представляющий объекты в двумерном пространстве, определяемом значениями некоторой пары переменных. В случае бóльшего числа переменных рассматриваются также матрицы графиков рассеивания. Однако в реальных исследовательских задачах количество признаков, регистрируемых на объектах наблюдения, часто может достигать сотни и даже более. В таких случаях наглядность представления близости объектов в пространстве признаков с помощью матрицы графиков рассеивания существенно снижается. Естественным желанием получить наглядное представление исходных данных, наиболее полно отражающее имеющуюся в них информацию, обусловлено появление методов редукции размерности признакового пространства, к которым относятся метод главных компонент, факторный анализ, многомерное шкалирование (МШ), корреспондентский анализ (КА, или анализ соответствий),

экстремальная группировка признаков и другие методы получения агрегированных показателей и отбора наиболее информативных признаков. Кроме необходимости наглядного представления данных применение методов редукции размерности может быть обусловлено желанием существенного сжатия объёмов хранимой статистической информации, а также стремлением к построению «лаконичных» моделей, упрощающих интерпретацию получаемых статистических выводов [15].

В процессе реализации процедур графических методов происходит построение некоторого пространства редуцированной размерности, в котором проявляют себя латентные факторы и становится очевидным действие этих факторов на пространственное расположение. В отличие от других статистических методов поиск координатного пространства осуществляется не по значениям конкретных признаков, а по данным представляющим сходства или различия этих объектов. На основе матрицы сходств / различий реализуется представление каждого объекта в виде точки геометрического пространства, координатами которой служат значения латентных факторов, в совокупности достаточно адекватно описывающих объект [79–81]. Таким образом, отношения между объектами становится удобным рассматривать в терминах расстояний как отношения между их точками-представителями [82]. Немаловажным преимуществом представления переменных в виде точек некоторого пространства является возможность измерения расстояния между ними. Это свойство графического представления использовано в работе при построении метода решения задачи классификации с обучением.

2.2.1 Многомерное шкалирование

Одним из методов, использованных в работе для обеспечения реализации первого этапа разработки информационной технологии, является многомерное шкалирование (МШ). Входными данными для техники МШ служат расстояния δ_{ij} между каждой парой объектов. Обратим внимание, что под понятием «объекты» здесь могут пониматься не только непосредственно объекты, составляющие обучающую выборку, но и признаки, их описывающие.

Расстояния между n объектами формируют $n \times n$-матрицу $D = \left(\delta_{ij}\right)_{i,j=1,..n}$.

Мы хотим представить имеющиеся n объектов в координатной системе более низкой размерности, но таким образом, чтобы в ней расстояния d_{ij} между этими объектами были максимально близки к исходным δ_{ij}: $\forall i, j = 1,..n:$ $\quad d_{ij} \cong \delta_{ij}$.

Наиболее часто окончательные расстояния d_{ij} измеряют в метрике Евклида, однако, в принципе, возможно применение и других метрик. Исходные расстояния δ_{ij} не всегда, собственно, являются именно расстояниями между объектами. Часто говорят о δ_{ij} как о мерах близости (или мерах различия), которые могут оцениваться, вообще говоря, не объективно с помощью каких-либо вычислительных процедур, а в результате экспертных суждений или других субъективных оценок.

Целью МШ является получение графической интерпретации, объясняющей, каким образом объекты связаны друг с другом, или

способной предоставить какую-либо другую содержательную интерпретацию данных.

Опишем процедуру метрического МШ, известного также в литературе как классическое решение (classical solution) и как анализ главных координат (principal coordinate analysis) [83]. Процедура была предложена У. Торгерсоном в [84].

На вход процедуры подаётся $n \times n$-матрица расстояний $D = (\delta_{ij})_{i,j=1,..n}$. С целью нахождения n точек в k-мерном пространстве ($k < n$) таких, что расстояния d_{ij} между ними приблизительно равны значениям δ_{ij}, выполняется следующая последовательность шагов:

1. Составляется $n \times n$-матрица A:

$$A = \left(\frac{-1}{2} \cdot \delta_{ij}^{\,2} \right)_{i,j=1,..n}.$$

2. Затем матрица A преобразовывается в симметричную матрицу B по формуле:

$$B = \left(I - \frac{1}{n} J \right) \cdot A \cdot \left(I - \frac{1}{n} J \right),$$

где I – единичная $n \times n$-матрица, J – $n \times n$-матрица, состоящая из единиц:

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & 1 & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}, \qquad J = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \dots & \dots & 1 & 1 \\ 1 & \dots & 1 & 1 \end{pmatrix}.$$

3. Матрица B представляется в виде своего спектрального разложения:

$$B = V \cdot \Lambda \cdot V^T,$$

где Λ – диагональная матрица собственных значений матрицы B, V – ортогональная матрица, состоящая из столбцов, являющихся собственными векторами матрицы B.

Если матрица B положительно полуопределённая, она имеет $q \le n$ положительных собственных значений. Оставшиеся $n - q$ собственных значений равны 0 (т.е. ранг матрицы $rank(B) = q \le n$): $\lambda_j > 0, \quad j = 1,..q; \ \lambda_j = 0, \quad j = q+1,...n$.

Таким образом, из-за возможности представления симметричной положительно полуопределённой $n \times n$-матрицы в виде её спектрального разложения, удаётся снизить размерность пространства до числа измерений, равного количеству собственных чисел матрицы B. Ортогональным базисом получаемого пространства являются собственные вектора $(v_1, \ \ldots, \ v_q)$ матрицы B. В этом базисе исходные n объектов, расстояния между которыми были заданы на входе, представляются n точками, координаты которых расположены по строкам $n \times q$-матрицы Z, определяемой соотношением:

$$Z = V_q \cdot \Lambda_q^{1/2} = \left(\sqrt{\lambda_1} \cdot v_1 \quad \sqrt{\lambda_2} \cdot v_2 \quad \ldots \quad \sqrt{\lambda_q} \cdot v_q\right).$$

При этом расстояния между этими n точками в q-мерном пространстве, натянутом на собственные векторы, совпадают с исходными расстояниями.

4. На практике количество собственных чисел q, как правило, получается достаточно большим, чтобы говорить о существенном снижении размерности пространства и, тем более, получить наглядное графическое представление данных. Поэтому выбирают первые (наибольшие) $k < q$ собственных значений и соответствующих им собственных

векторов, таким образом, чтобы расстояния d_{ij} между n точками в этом k-мерном пространстве достаточно удовлетворительно приближали исходные δ_{ij}.

Наиболее часто используют проекцию в двумерное пространство ($k = 2$). Она наиболее легка для графического представления, наглядности восприятия и, соответственно, для интерпретации (рис. 2.2).

Рис. 2.2. Пример карты, представляющей 10 парных взаимосвязей между пятью признаками в двумерном пространстве, полученной методом многомерного шкалирования

Для построения шкального пространства и интерпретации порядковых данных применяют методы неметрического МШ. Впервые метод неметрического МШ был предложен Р. Шепардом в [85]. В неметрическом МШ предполагается, что $m = n(n-1)/2$ расстояний (в более общем смысле – мер различия) δ_{ij} $(i, j = 1,..n)$ между n объектами не могут быть измерены непосредственно и

оценены численно, однако могут быть ранжированы в каком-либо порядке: $\delta_{r_1 s_1} < \delta_{r_2 s_2} < \ldots < \delta_{r_m s_m}$.

Задача неметрического МШ состоит в нахождении пространства сниженной размерности, в котором сохраняется монотонность связей между объектами. Другими словами, в окончательном пространстве исходные *n* объектов представляются точками, расстояния d_{ij} между которыми должны быть упорядочены в той же последовательности, что и исходные расстояния: $d_{r_1 s_1} < d_{r_2 s_2} < \ldots < d_{r_m s_m}$.

Вид монотонности заранее неизвестен, и функция, наилучшим образом приближающая упорядоченность исходных данных, подбирается эмпирическим путём. Наиболее часто используют линейную, степенную, показательную или логарифмическую зависимость [86].

Процедура неметрического МШ состоит в последовательной реализации следующего ряда шагов:

1. Сначала выбирается размерность *k* пространства окончательной конфигурации.
2. Строится некоторая стартовая конфигурация из *n* точек в пространстве выбранной размерности *k*.

Построение начальной конфигурации может быть осуществлено на основе различных подходов. Например, для этого можно использовать процедуру метрического МШ, относясь к порядковым мерам близости как к метрическим (интервальным) величинам [77]. Можно использовать метод простой ординации Орлочи [87], алгоритм Торгерсона [88], Краскала [89] или др. Однако, *n* точек, расстояния между которыми монотонно возрастают в пространстве выбранной размерности *k*, могут быть выбраны и как случайные числа из

равномерно или нормально распределённой генеральной совокупности [77, 86].

3. В полученной конфигурации проверяется соответствие исходных расстояний воспроизведённым. Затем начальная конфигурация улучшается.

При выполнении алгоритма поиска оптимального шкального пространства минимизируется мера несоответствия между исходными расстояниями и расстояниями, воспроизведёнными в пространстве заданной размерности:

$$\sum_{i,j}\left(\delta_{ij} - f\left(d_{ij}\right)\right)^2 \to \min,$$ где f – некоторая монотонная функция.

Качество полученной проекции (конфигурации) оценивается по величине стресса, формула вычисления и содержательная интерпретация которой предложены Краскалом в [90]:

$$S = \sqrt{\sum_{i<j}\left(\delta_{ij} - d_{ij}\right)^2 \Big/ \sum_{i<j}\delta_{ij}^{\,2}}. \qquad (2.1)$$

Значения стресса более 20% говорят о плохом качестве конфигурации, от 10% до 20% – об удовлетворительном; 5—10% стресса соответствуют хорошему качеству проекции, до 2,5% – отличному.

Другая формула для оценки величины стресса вводится Краскалом и Вишем в [91]:

$$\varphi = \sqrt{\sum_{i,j}\left(d_{ij} - \delta_{ij}\right)^2 \Big/ \sum_{i,j}\left(d_{ij} - \bar{d}\right)^2},$$

где $\bar{d}$ – среднее арифметическое всех оцененных расстояний.

Существуют и другие модификации формул для нахождения стресса (Юнг):

$$S = \sqrt{\sum_{i,j}\left(\delta_{ij}{}^2 - d_{ij}{}^2\right)^2 \Big/ \sum_{i,j} d_{ij}{}^4}\ ;$$

$$\varphi = \sqrt{\sum_{i,j}\left(d_{ij}{}^2 - \delta_{ij}{}^2\right)^2 \Big/ \sum_{i,j}\left(d_{ij}{}^2 - \bar{d}^2\right)^2}\ .$$

Льюисом Гуттманом в [92] также была предложена другая характеристика качества окончательной конфигурации – коэффициент отчуждения (alienation):

$$k = \left(1 - \left(\sum_{i,j} d_{ij}\delta_{ij}\right)^2 \Big/ \left(\sum_{i,j} d_{ij} \cdot \sum_{i,j} \delta_{ij}\right)\right)^{1/2}. \qquad (2.2)$$

В данной работе для исследования взаимосвязей между целевым и объясняющими признаками предлагается рассматривать карты, полученные с помощью процедур неметрического МШ. Наилучшее представление выбирать на основании значений коэффициентов стресса (2.1) и отчуждения (2.2).

2.2.2 Корреспондентский анализ

Второй из методов графического представления многомерных данных, использованных для решения задач этапа 1 разработки информационной технологии, – метод корреспондентского анализа (КА, или, анализа соответствий). Простой КА предназначен для графического отображения связей между категориями переменных в двухвходовой таблице сопряжённости. Входными данными для простого КА является результат кросстабуляции двух качественных переменных.

Пусть a и b – количество категорий каждой из кросстабулированных переменных. В таблице сопряжённости

(рис. 2.3) показаны абсолютные частоты совместного появления категорий каждой из переменных n_{ij} $(i=1,2,..a; j=1,2,...b)$.

i \ j	1	2	…	b
1	n_{11}	n_{12}	…	n_{1b}
2	n_{21}	n_{22}	…	n_{2b}
…	…	…	…	…
a	n_{a1}	n_{a2}	…	n_{ab}

Рис. 2.3. Двухвходовая $a \times b$-таблица сопряжённости качественных переменных

Суммы по строкам и столбцам данной таблицы называются маргинальными частотами и обозначаются: $n_{i+} = \sum_{j=1}^{b} n_{ij}$,

$n_{+j} = \sum_{i=1}^{a} n_{ij}$. Общее число наблюдений

$$n = \sum_{i=1}^{a} n_{i+} = \sum_{j=1}^{b} n_{+j} = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij}.$$

По таблице сопряжённости вычисляются относительные частоты p_{ij}, из которых формируется т.н. матрица соответствий (correspondence matrix) P:

$$P = \left(p_{ij}\right)_{i=1,..a}^{j=1,..b}, \qquad p_{ij} = n_{ij} / n.$$

Суммы относительных частот по строкам матрицы P в КА принято называть массами строк и обозначать:

$$p_{i+} = \sum_{j=1}^{b} p_{ij} = \sum_{j=1}^{b} n_{ij} / n.$$

Аналогично, суммы относительных частот по столбцам матрицы соответствий называют массами столбцов и обозначают:

$$p_{+j} = \sum_{i=1}^{a} p_{ij} = \sum_{i=1}^{a} n_{ij} \Big/ n .$$

Профили строк и столбцов получаются в результате деления элементов матрицы соответствий на массу соответствующей строки или соответствующего столбца. Так, профиль i-ой строки – это $1\times b$-вектор-строка

$$\left(\frac{p_{i1}}{p_{i+}} \quad \frac{p_{i2}}{p_{i+}} \quad \cdots \quad \frac{p_{ib}}{p_{i+}} \right).$$

Профиль j-го столбца – это $a\times 1$-вектор-столбец:

$$\left(\frac{p_{1j}}{p_{+j}} \quad \frac{p_{2j}}{p_{+j}} \quad \cdots \quad \frac{p_{aj}}{p_{+j}} \right)^{\mathrm{T}}.$$

Обозначим через r $a\times 1$-вектор-столбец, компонентами которого являются массы строк; через c $b\times 1$-вектор-столбец, состоящий из масс столбцов:

$$r = \left(p_{1+} \quad p_{2+} \quad \cdots \quad p_{a+}\right)^{\mathrm{T}}, \qquad c = \left(p_{+1} \quad p_{+2} \quad \cdots \quad p_{+b}\right)^{\mathrm{T}}.$$

Введём матрицы:

$$D_r = diag(r) = \begin{pmatrix} p_{1+} & 0 & \cdots & 0 \\ 0 & p_{2+} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & p_{a+} \end{pmatrix},$$

$$D_c = diag(c) = \begin{pmatrix} p_{+1} & 0 & \cdots & 0 \\ 0 & p_{+2} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & p_{+b} \end{pmatrix}.$$

Далее строится $a \times b$-матрица $G = D_r^{-1/2} \cdot \left(P - r \cdot c^{\mathrm{T}}\right) \cdot D_c^{-1/2}$. Другими словами, элементы матрицы $G = \left(g_{ij}\right)_{\substack{i=1,..a \\ j=1,..b}}$ определяются по формуле:

$$g_{ij} = \frac{\left(p_{ij} - p_{i+} \cdot p_{+j}\right)}{\sqrt{p_{i+} \cdot p_{+j}}}, \quad \forall i = 1,..a, \quad \forall j = 1,..b.$$

По сути, элементы матрицы G представляют собой стандартизированные отклонения строк и столбцов от независимости. Действительно, для проверки гипотезы о независимости качественных переменных, кросстабулированных в таблице сопряжённости, используется статистика $\chi 2$ Пирсона [93—95]:

$$\chi 2 = n \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{\left(p_{ij} - p_{i+} \cdot p_{+j}\right)^2}{p_{i+} \cdot p_{+j}},$$

которая с учётом введенных выше обозначений может быть записана в векторной форме следующим образом:

$$\chi 2 = n \cdot tr\left[D_r^{-1} \cdot \left(P - r \cdot c^{\mathrm{T}}\right) \cdot D_c^{-1} \cdot \left(P - r \cdot c^{\mathrm{T}}\right)^{\mathrm{T}} \right] = n \cdot tr\left[G \cdot G^{\mathrm{T}}\right].$$

Находится сингулярное разложение матрицы G в виде:

$$G = U \cdot \Lambda \cdot V^{\mathrm{T}},$$

где U – $a \times q$-матрица левых сингулярных векторов, т.е. собственных векторов матрицы $G \cdot G^{\mathrm{T}}$;

V – $q \times b$-матрица правых сингулярных векторов, т.е. собственных векторов матрицы $G^{\mathrm{T}} \cdot G$;

$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_q \end{pmatrix}$ – диагональная $q \times q$-матрица сингулярных значений G,

т.е. $\lambda_1^2, \lambda_2^2, \ldots \lambda_q^2$ – ненулевые собственные значения матрицы $G \cdot G^{\mathrm{T}}$ (или $G^{\mathrm{T}} \cdot G$).

Количество ненулевых собственных значений q равно:

$$q = \operatorname{rank}\left(G \cdot G^{\mathrm{T}}\right) = \operatorname{rank}(G) = \operatorname{rank}\left(P - r \cdot c^{\mathrm{T}}\right) = \min\{a-1; b-1\}.$$

Матрица U является ортонормированным базисом, в котором находятся координаты точек, представляющих строки таблицы сопряжённости, по формуле:

$$X = D_r^{-1/2} \cdot U \cdot \Lambda .$$

Т.е. s-я координата $(s = 1, \ldots q)$ i-ой строки $(s = 1, \ldots q)$ определяется по формуле:

$$x_{is} = \frac{\lambda_s}{\sqrt{p_{i+}}} \cdot u_{is} .$$

Столбцы матрицы V представляют собой ортонормированный базис, в котором находятся координаты точек, представляющих столбцы таблицы сопряжённости. Для этого используется формула:

$$Y = D_c^{-1/2} \cdot V \cdot \Lambda .$$

Т.е. s-я координата $(s = 1, \ldots q)$ j-го столбца $(j = 1, \ldots b)$ определяется по формуле:

$$y_{js} = \frac{\lambda_s}{\sqrt{p_{+j}}} \cdot v_{js} .$$

В общем случае матрица G не является симметрической, поэтому базисы U и V не совпадают. Однако, за счёт того, что собственные числа матриц $G \cdot G^{\mathrm{T}}$ и $G^{\mathrm{T}} \cdot G$ одни и те же, и их количество одинаково (равно q), становится возможным представить точки, отвечающие строкам, и точки, отвечающие столбцам, на одной карте. При этом в большинстве работ, касающихся методики простого КА, делается ударение на то, что расстояния между точками, отвечающими строкам таблицы сопряжённости, также как и расстояния между точками, отвечающими столбцам таблицы сопряжённости, отражают силу причинно-следственной связи между категориями одной из переменных. В то же время расстояние между точкой-строкой и точкой-столбцом подобного смысла не имеет. Т.е. близость точки-строки и точки-столбца на карте соответствий в общем случае ничего не говорит о причинно-следственной связи между ними, наблюдаемой по эмпирическим данным, представленным таблицей сопряжённости.

В [96, 77, 97, 15] предлагается оценивать близость между категориями разных переменных на основании угла между векторами, соединяющими начало координат с точками, их представляющими, подобно подходу, существующему в анализе главных компонент. В [98] предлагается построение «несимметричной карты», на которой становится возможным правильно трактовать расстояние между точкой-строкой и точкой-столбцом.

Множественный КА является обобщением простого КА на случай многовходовых таблиц сопряжённости, кросстабулирующих $l > 2$ качественных переменных. При его выполнении на основании многовходовой таблицы сопряжённости строится т.н. индикаторная

матрица I. Если m_i – количество категорий i-ой качественной переменной, участвующей в анализе, $(i = 1,\ldots l)$; $m = \sum_{i=1}^{l} m_i$ – общее количество категорий всех переменных в многовходовой таблице сопряжённости, то размерность индикаторной матрицы будет равна $n \times m$, где n – количество наблюдений. Индикаторная матрица I состоит только из нулей и единиц. Если наблюдение принадлежит некоторой категории, то элемент на пересечении соответствующих строки и столбца равен 1, в противном случае – 0.

На основании индикаторной матрицы составляется матрица Бёрта (B), впервые использованная в [99]. Это квадратная $m \times m$-матрица, кросстабулирующая связи между всеми имеющимися переменными и представляющая собой матричное произведение транспонированной и исходной индикаторной матрицы:

$$B = I^{\mathrm{T}} \cdot I.$$

Матрица Бёрта симметрична, а суммы диагональных элементов в каждом блоке, представляющем кросстабуляцию некоторой переменной с самой собой, равны общему числу объектов. Внедиагональные блоки являются двухвходовыми таблицами сопряжённости i-ой и j-ой переменных $(i, j = 1,\ldots l, i \neq j)$.

Процедура КА выполняется над матрицей Бёрта и по описанной выше методике получаются координаты точек, отвечающих каждой из категорий входных переменных.

За счёт симметричности матрицы B категории разных переменных удаётся представить точками не только на одной карте, но и в едином координатном пространстве (т.е. в одном базисе). В силу этого становится возможным измерять расстояния между

точками, отвечающими категориям различных качественных переменных, и близость этих точек можно трактовать как наличие причинно-следственных связей между определёнными категориями различных переменных. Это свойство графического представления, получаемого методом множественного КА, использовано в данной работе при построении метода классификации на основании геометрической интерпретации структуры данных (параграф 3.2).

Общей инерцией графической интерпретации, полученной методами КА, называется величина χ^2 / n, которую в силу приведенных выше формульных выкладок можно представить в виде:

$$\frac{\chi^2}{n} = \sum_{i=1}^{q} \lambda_i^2 ,$$

где q – количество собственных чисел (или, ранг) матрицы G.

На практике, как и в МШ, для представления проекции (карты) в корреспондентском анализе, как правило, выбирается число измерений $k < q$. Качество представления связей между категориями переменных, кросстабулированных в исходной таблице сопряжённости, на k-мерной карте принято оценивать по величине инерции, которая для k выбранных измерений представления вычисляется по формуле [97, 77, 96]:

$$\mathrm{I} = \sum_{i=1}^{k} \lambda_i^2 \Big/ \sum_{i=1}^{q} \lambda_i^2 \cdot 100\% .$$

При этом процент инерции, объяснённой каждым s-м измерением $(s = 1,2,..k)$, находится как $\dfrac{\lambda_s^2}{\sum_{i=1}^{q} \lambda_i^2}$.

Применение техники множественного КА в данной работе, позволило получать графические представления классов и категоризированных объясняющих переменных, которые обеспечивали решение задач первого этапа построения информационной технологии (разработки метода классификации с обучением).

2.2.3 Методы выбора оптимальной размерности пространства проекции

Формально и с точки зрения чисто математического подхода, для выбора размерности пространства окончательной конфигурации можно использовать произвольное число измерений [90]. Выбор размерности определяется целями исследования, и/или осуществляется в результате численных экспериментов. Вообще, как говорится в [90], это «дело суждения учёного-исследователя» ("*it is a matter of scientific judgment*"). Однако существуют и некоторые более или менее общепринятые подходы к решению этого вопроса.

Для выбора оптимальной размерности пространства отображения в методах графической интерпретации структуры данных часто используется критерий Кэттеля, который изначально был предложен для использования в факторном анализе [100]. Критерий основан на рассмотрении т.н. «графика каменистой осыпи» (scree plot). График отражает убывание вклада в общее качество графической модели каждого нового измерения. В МШ на графике каменистой осыпи принято изображать величину стресса, в КА – долю инерции, объяснённой каждым измерением. Критерий состоит в поиске точки, где убывание собственных значений замедляется

наиболее сильно. В [32] данный критерий назван «критерием отсеивания». Следует отметить, что критерий не является строго математически обоснованным, поэтому иногда целесообразно использовать бóльшее количество измерений для достижения лучшего качества проекции.

На практике ещё иногда выбирают размерность, соответствующую количеству собственных чисел, которые больше определённого порогового значения. На этом подходе основаны критерий Кайзера [101] (измерения, соответствующие собственным числам менее 1, считаются малоинформативными; применяется в факторном анализе) и критерий Жоффе [102] (отбираются собственные числа больше 0,7).

Наиболее полно существующие подходы к определению оптимального количества измерений описаны в [32]. Хотя это описание и касается выбора оптимального числа факторов для процедуры факторного анализа, основные его принципы перенесены и на выбор числа объясняющих осей координатного пространства в методах МШ и КА. Наиболее обоснованными на наш взгляд являются критерии, основанные на интерпретируемости и инвариантности модели [103, 104, 90, 32]. Т.е. полученные результаты оцениваются, во-первых, с точки зрения приемлемости и ясности модели в данной предметной области. Во-вторых, если при применении различных критериев оценки качества полученной модели и их комбинаций, получаются одинаковые или схожие результаты, это служит показателем её адекватности.

В данной работе для выбора размерности пространства геометрического представления наряду с критерием Кэттеля применялись подходы, основанные на инвариантности модели, однако

основным критерием служило качество получаемой проекции. При использовании МШ выбирались представления с наилучшими значениями стресса и отчуждения, в случае применения КА – модели, объясняющие бóльший процент инерции.

2.3 Методы поиска признаков, определяющих различия между классами

Для решения задачи предварительного этапа (этап 0, рис. 2.1) разработки информационной технологии необходимо было сформировать набор входных признаков для модели классификатора, выбрав из общего количества только те показатели, которые будут полезными для объяснения правил отнесения объектов к классам. При выборе информативных (в смысле, определяющих принадлежность объектов к классам) признаков в данной работе мы исходили из соображений значимости влияния конкретного признака на принадлежность объектов к классам, которая оценивалась на основании статистически значимого различия между значениями признака в разных классах. Т.к. априори исходная обучающая информация может состоять из разнородных признаков, то и для оценки значимости их различий в классах возникла необходимость использования разных статистических методов, обусловленная типом данных, шкалой измерения и видом закона распределения переменных.

Для количественных признаков наиболее известным методом определения значимости различия значений в нескольких группах является дисперсионный анализ (в случае двух классов (групп) – Т-критерий Стьюдента). Но так как применение этих критериев

ограничено довольно строгими требованиями, накладываемыми на исходные данные (нормальность распределения и однородность дисперсий показателей в классах, которые в случае экспериментальных данных выполняются одновременно только в около 4-5% (или даже менее) случаев [105, 106, 30]), то для достижения поставленной задачи в работе использован ряд непараметрических аналогов этих методов.

Для выявления статистически значимых отличий значений признака в нескольких классах использовался анализ Краскала—Уоллиса [107], предназначенный для проверки нулевой гипотезы о значимости различия параметров сдвига нескольких (более двух) групп [93].

При множественных попарных сравнениях между классами применялись непараметрические критерии сравнения двух выборок (Вилкоксона—Манна—Уитни [108, 109], Вальда—Вольфовица [110], Колмогорова—Смирнова [111–114]). При проведении тестов парных сравнений классов учитывалась поправка на множественность в формулировке Бонферрони [115]. Её суть в том, что при проведении серии парных сравнений среди $k > 2$ групп для уменьшения вероятности случайного ошибочного отвержения нулевой гипотезы предельно допустимый уровень значимости каждого парного сравнения корректируется по формуле: $\alpha' = \alpha / m$, где m – число проведенных сравнений. Данный подход к решению проблемы множественных сравнений, является самым простым, но считается многими исследователями достаточно жёстким и неэффективным. Считается, что при нём уровень значимости слишком занижается, что ведёт к возрастанию вероятности ошибки второго рода, т.е. снижению

статистической мощности исследования. В этом случае, особенно при достаточно большом количестве сравниваемых групп, действительно значимые различия могут быть не обнаружены. Однако, если число сравнений невелико (не больше 8), то результаты, полученные при использовании метода Бонферрони, существенно не отличаются от результатов при использовании других улучшенных поправок на множественность. Это послужило обоснованием целесообразности применения данного метода при решении практической задачи настоящего исследования.

Влияние качественных признаков на принадлежность к классам оценивалось с помощью анализа $a \times b$–таблиц сопряжённости показателя «класс» с этими признаками. Для анализа в зависимости от распределения частот в клетках таблицы сопряжённости использовались точный тест Фишера, критерий $\chi 2$, или критерий $\chi 2$ с поправкой Йетса [116, 117, 94, 95].

2.4 Алгоритмы вычисления оценок

Алгоритм вычисления оценок (АВО) использован в работе на этапе 2 разработки информационной технологии для формирования решающих правил отнесения объектов к классам. Модели АВО относятся к классу моделей, основанных на частичной прецедентности, так как базируются на анализе сходства между признаковыми описаниями ранее классифицированных объектов и объектов, которые необходимо распознать. Этот класс моделей известен также в литературе под названием моделей голосования, или Г-моделей [118, 14, 119]. Модель АВО считается одной из базовых

моделей алгоритмов распознавания и классификации [120, 121], которую можно использовать в качестве языка описания методов распознавания [122], а в [18, 123] эту модель называют канонической моделью классификатора.

Принцип действия моделей АВО состоит в последовательной реализации двух этапов. На первом этапе распознающий оператор строит вектор оценок, характеризующих принадлежность объекта к классам. В классической модели эти оценки вычисляются на основании сходства классифицируемого объекта с некоторым эталонным образцом, которое определяется по некоторой системе подмножеств признакового описания двух объектов и конкретизируется выбором определённой функции близости. На втором этапе решающее правило на основании вектора оценок вырабатывает решение об отнесении объекта к тому или иному классу. Как правило, объект относится к тому классу, для которого получена максимальная оценка [124–126, 121, 118, 14, 119]. Поэтому модели АВО часто представляются в виде суперпозиции двух функций – распознающего оператора b и решающего правила r [125, 127]:

$$A(S) = r\big(b(S)\big),$$

где S – объект, который необходимо классифицировать.

Тем, каким образом будет задаваться распознающий оператор, и каким образом будет действовать решающее правило, и определяется модель конкретного алгоритма вычисления оценок. В наиболее обобщённом виде этапы, необходимые для идентификации модели АВО, описаны в [118, 14]. Здесь мы приведём их сжато, укрупнив и объединив некоторые шаги этого процесса.

Для начала уточним некоторые обозначения. Пусть решается задача классификации на l классов $\{C_j\}_{j=1}^{l}$. Обучающая информация задана в виде $m \times n$-таблицы обучения. $\{S_u\}_{u=1}^{m}$ – объекты обучающей выборки (строки матрицы обучения); $\{a_k\}_{k=1}^{n}$ – признаки, описывающие объекты (столбцы матрицы обучения).

На первом этапе идентификации модели АВО выбирается система опорных множеств Ω_A алгоритма A, представляющая собой некоторую систему подмножеств множества $\{1, \ 2, \ \ldots \ n\}$. По сути, указанием системы опорных множеств определяются наборы признаков (точнее, номера столбцов, их представляющих), которые будут входить в модель АВО.

Второй этап состоит в выборе функции, с помощью которой будет оцениваться расстояние между объектами, – функции близости $B_\Omega(S_u, S_t)$. Индекс Ω указывает здесь на то, что близости между объектами вычисляются по опорным множествам алгоритма, т.е. по поднаборам (т.н. Ω-частям) их признакового описания.

Третий этап идентификации модели АВО предусматривает определение оценок $\{\Gamma_j(S)\}_{j=1}^{l}$ объекта S по классам $\{C_j\}_{j=1}^{l}$. Эти оценки обычно определяются через суммирование или усреднение оценок принадлежности объекта к классам по элементам системы опорных множеств $\Omega \in \Omega_A$:

$$\Gamma_j(S) = \sum_{\Omega \in \Omega_A} \Gamma_\Omega^j(S) \qquad \text{или} \qquad \Gamma_j(S) = \frac{1}{N} \cdot \sum_{\Omega \in \Omega_A} \Gamma_\Omega^j(S),$$

где N – нормирующий множитель (количество слагаемых в сумме);

$\Gamma^j_\Omega(S)$ – оценка принадлежности объекта S к классу C_j по опорному множеству $\Omega \in \Omega_A$, являющаяся функцией от близостей объекта ко всем объектам класса C_j по опорному множеству Ω:

$$\Gamma^j_\Omega(S) = f\left(B_\Omega(S,S_{u_1}),\ldots,B_\Omega(S,S_{u_t}),\quad p(\Omega),\quad \gamma(S_{u_1}),\ldots,\gamma(S_{u_t})\right),$$

где $\{S_{u_1},\ldots,S_{u_t}\} \subset \{S_1,\ldots,S_m\}$, таких, что $S_{u_i} \in C_j \quad \forall i = 1,2,\ldots,t$, т.е. множество объектов обучающей выборки, относящихся к классу C_j;

$p(\Omega)$ – вектор числовых параметров, называемый весами признаков, определяемых опорным множеством $\Omega \in \Omega_A$;

$\left(\gamma(S_{u_1}),\ \ldots,\ \gamma(S_{u_t})\right)$ – вектор числовых параметров, называемый весами объектов $S_{u_1},\ldots,S_{u_t}$.

И, наконец, на четвёртом этапе указывается вид решающего правила, которое на основании оценок объекта по классам относит его к одному из них. Как уже говорилось выше, наиболее часто употребляемым решающим правилом является операция выбора максимального элемента множества оценок: $\max\limits_{j=1,l} \Gamma_j(S)$.

Наиболее общий вид решающего правила может быть описан следующей формулировкой. Для каждого класса C_j ($j = 1, \ldots, l$) выбирается функция f_j (на практике обычно из семейства линейных функций) и две пороговых константы $c_{1j} < c_{2j}$. Решение об отнесении объекта S к классу C_j выносится по правилу:

$$r(S) = \begin{cases} S \in C_j, & \text{если} \quad f_j\left(\Gamma_1(S),\ldots,\Gamma_l(S)\right) > c_{2j} \\ S \notin C_j, & \text{если} \quad f_j\left(\Gamma_1(S),\ldots,\Gamma_l(S)\right) < c_{1j} \\ S \in \emptyset, & \text{если} \quad c_{1j} \le f_j\left(\Gamma_1(S),\ldots,\Gamma_l(S)\right) \le c_{2j} \end{cases} \qquad \forall j = 1,\ldots,l.$$

(Последний вариант соответствует отказу от классификации объекта.)

Основные характеристики модели АВО, определяемые на каждом из описанных выше четырёх этапов её идентификации, показаны на рис. 2.4. Выбором различных правил определения параметров (1)—(4) и порождается всё многообразие возможных моделей алгоритмов вычисления оценок.

В настоящей работе предложен метод построения классификаторов по обучающей информации в виде моделей АВО со следующими характеристиками. Система опорных множеств алгоритма определяется на предварительном этапе (этапе 0, рис. 2.1) разработки информационной технологии. Этап 1 определяет способы оценки близости объектов и нахождения оценок по классам, которые осуществляются на основании геометрической интерпретации структуры данных. Решающее правило модели АВО, использованное в работе на этапе 2 разработки информационной технологии, – операция выбора максимальной оценки.

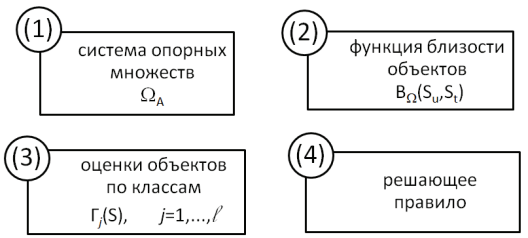


Рис. 2.4. Основные характеристики, идентифицирующие модель АВО

2.5 Методы построения композиций классификаторов

Наличие этапа 3 (рис. 2.1) решения основной задачи данной работы было обусловлено необходимостью повышения точности

классификации, которое достигалось за счёт построения композиций, включающих несколько моделей классификаторов. Подобные комбинации, составленные из нескольких моделей, называются в литературе алгоритмическими композициями [127–129], комитетами алгоритмов [130–133] или ансамблями [134–136]. Их формирование считается наиболее сильным приёмом в случаях, когда при решении практических задач ни один из стандартных подходов (выбор модели из другого семейства алгоритмов, или использование альтернативных методов настройки её параметров) не позволяет получить классификатор приемлемой точности, либо если различными методами удаётся построить несколько удачных классификаторов с соизмеримыми характеристиками качества [13, 15, 127, 129, 137]. В работах [130, 138, 134, 18, 137] комбинирование моделей признано наиболее перспективным направлением для повышения точности классификации и распознавания.

Для построения композиций классификаторов разработан ряд методов, однако все они в той или иной мере используют две основных идеи: специализация и взвешенное голосование.

2.5.1 Композиции на основе взвешенного голосования

Простое голосование, или голосование по большинству (majority vote) [139–144] было исторически первым методом построения композиций. При использовании простого голосования по большинству результирующий ответ комитета выбирается на основе большинства ответов, входящих в композицию базовых алгоритмов. В задачах классификации это означает, что классифицируемый образец

относится к тому классу, к которому его отнесли большинство входящих в комитет классификаторов.

Обозначим $\{T^k\}_{k=1}^m$ – множество из m классификаторов, решающих задачу дискриминации объектов $x \in X$ на n классов $\{C_i\}_{i=1}^n$. Ответ базового алгоритма T^k на некотором объекте x: $T^k(x) = C_i$, где $i = 0, 1, \ldots, n$. При этом под $T^k(x) = C_0$ подразумевается отказ алгоритма T^k от классификации объекта x.

Если ответы базовых классификаторов задать в виде n-мерных бинарных векторов [18]:

$$[d_{k1},\ldots,d_{kn}]^{\mathrm{T}} \in \{0;1\}^n \quad (k = 1,\ldots,m), \qquad \text{где}$$

$$d_{ki} = \begin{cases} 1, & \text{если} \quad T^k(x) = C_i \\ 0, & \text{если} \quad T^k(x) \neq C_i \end{cases}.$$

Тогда суммы $\sum_{k=1}^m d_{ki}$ будут означать количество голосов, отданных за каждый из классов C_i $(i = 1,\ldots,n)$. Следовательно, ответ композиции классификаторов, составленной по методу голосования по большинству, можно формально представить следующим образом [18]:

$$T(x) = C_\theta, \qquad \text{где} \qquad \theta = \arg \max_{i=1}^{n} \sum_{k=1}^{m} d_{ki}.$$

Другое формульное выражение для композиции голосования по большинству находим в [127, 145]:

$$T(x) = \frac{1}{m} \sum_{k=1}^{m} T^k(x).$$

Как усовершенствование метода голосования по большинству

появилось взвешенное голосование (weighted majority vote) [139, 146, 18, 147]. В этом методе голоса базовых моделей учитываются с весами, как правило, зависящими от их точности на обучающей выборке. В [127] даётся следующее формальное представление композиции на основе взвешенного голосования:

$$T(x) = \sum_{k=1}^{m} \alpha_k \cdot T^k(x),$$

где α_k – весовые коэффициенты базовых классификаторов ($k = 1, \ldots, m$).

Согласно [18] можно представить ответ композиции на основе взвешенного голосования более строгим определением:

$$T(x) = C_\theta, \qquad \text{где} \qquad \theta = \arg \max_{i=1}^{n} \sum_{k=1}^{m} \alpha_k \cdot d_{ki}.$$

2.5.2 Композиции, основанные на принципе специализации

Принцип специализации основан на манипулировании обучающим множеством с последующим построением базовых классификаторов на различных его подмножествах [148]. Композиции, построенные на основе принципа специализации, считаются особенно эффективными в ситуациях, когда базовые алгоритмы являются неустойчивыми к небольшим изменениям в обучающей информации. На основе принципа специализации построен метод голосования по старшинству, известный также под названиями списка решающих правил или машины покрывающих множеств (Set Covering Machine, SCM) [133, 149, 150].

При голосовании по старшинству классифицируемый объект

$x \in X$ последовательно передаётся от одного базового алгоритма T^k (k =1, …, m) к другому до тех пор, пока хотя бы один из них не выдаст ответ $T^k(x) = C_k$, т.е. «покроет» объект. В случае, если ни один из алгоритмов не покрыл объект x, комитет отказывается от классификации: $T(x) = C_0$.

При формировании композиции классификаторов с помощью голосования по старшинству возникает проблема выбора порядка функционирования базовых алгоритмов, или, что эквивалентно, формирования последовательности классов, на которых один за другим будут выдавать ответы базовые алгоритмы [127, 151]. В случае некорректного определения порядка классов композиция будет выдавать неприемлемые решения, и потенциальное преимущество комбинирования классификаторов будет потеряно. В [127] предложены стратегии упорядоченности последовательности классов, причём указывается, что выбор конкретной стратегии из предложенных обуславливается особенностями решаемой прикладной задачи. Приоритетный порядок классов можно выбирать на основе:

1) лучшего значения качества распознавания;

2) количества непокрытых объектов в классе;

3) порядок может задаваться предварительно (навязываться), а затем базовые классификаторы строятся как одноклассовые, настроенные на отделение объектов только одного данного класса от объектов всех остальных классов.

Последний способ построения композиций голосования по старшинству считается наиболее легко поддающимся содержательной интерпретации, а потому является наиболее часто используемым.

Принцип специализации послужил основой для создания и такого способа построения композиций, как смеси экспертов. Архитектура смеси экспертов (рис. 2.5 [18]) была изначально предложена для нейронных сетей [152–155]. Экспертами (базовыми алгоритмами) являются нейронные сети, настроенные таким образом, что каждая из них отвечает за определённую часть признакового пространства. Селектор использует выходной результат (ответ) другой нейронной сети, называемой шлюзовой сетью. На вход шлюзовой сети поступает классифицируемый объект $x \in X$, а выходом является набор коэффициентов $\alpha_1(x), ..., \alpha_m(x)$, называемых шлюзами (gates) [152, 18] или функциями компетентности [127, 146]. Как правило, коэффициенты нормируются, т.е. $\sum_{k=1}^{m} \alpha_k(x) = 1$, и к ним предъявляется требование неотрицательности: $\alpha_k(x) \geq 0$ ($\forall\ k = 1, \ldots, m$). При этом каждый $\alpha_k(x)$ интерпретируется как вероятность того, что эксперт T^k наиболее компетентный эксперт для правильной классификации конкретного объекта x.

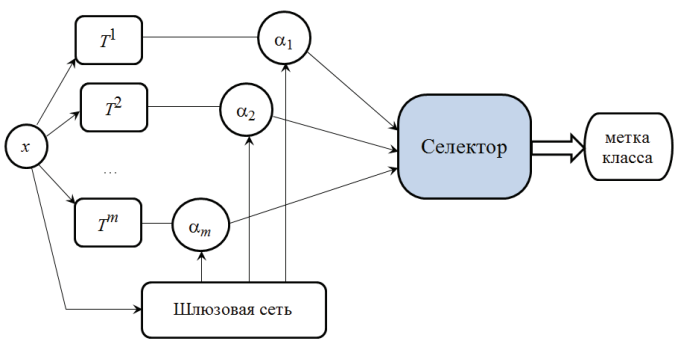


Рис. 2.5. Алгоритм функционирования смеси экспертов [18]

В механизме действия селектора вероятности $\alpha_k(x)$ и ответы базовых классификаторов на объекте x собираются вместе для получения окончательного ответа композиции одним из следующих способов:

1) «Победитель забирает всё» [18, 151], когда результирующий ответ композиции соответствует тому классу C_i, вес которого на данном объекте $\alpha_i(x)$ максимальный:

$$T(x) = C_\theta, \qquad \text{где} \qquad \theta = \arg \max_{k=1}^{m} \alpha_k(x).$$

2) Стохастический выбор, при котором «главный» классификатор, ответ которого принимается за результирующий ответ композиции, выбирается на основании распределения $\alpha_1(x)$, ..., $\alpha_m(x)$.

3) Взвешенный выбор, при котором результирующий ответ композиции:

$$T(x) = C_\theta, \qquad \text{где} \qquad \theta = \arg \max_{i=1}^{n} \sum_{k=1}^{m} \alpha_k(x) \cdot d_{ki}.$$

Взвешенный выбор является наиболее часто используемым способом функционирования селектора, поэтому формально смесь экспертов стандартно представляется в виде [127]:

$$T(x) = \sum_{k=1}^{m} \alpha_k(x) \cdot T^k(x).$$

Если функция $\alpha_k(x)$ принимает только два значения {0, 1}, то множество всех $x \in X$, для которых $\alpha_k(x) = 1$, называется областью компетентности [146] базового алгоритма T^k. В общем случае функция $\alpha_k(x)$ описывает область компетентности как нечёткое

множество, и значение $\alpha_k(x) \in [0, 1]$ рассматривается как степень принадлежности объекта x области компетентности базового алгоритма T^k.

В настоящей работе для разработки метода построения композиций классификаторов “рейтинговое голосование” (параграф 3.2.1) использована комбинация эвристических принципов, лежащих в основе взвешенного голосования и смесей экспертов; для разработки метода построения композиций классификаторов “рейтинговое голосование по старшинству” (параграф 3.2.2) к этим эвристикам добавлен принцип действия комитета с логикой старшинства.

Исследования, результаты которых изложены в данном разделе, позволяют сформулировать следующие выводы.

1. Анализ методов графической интерпретации структуры данных позволил выделить метод множественного корреспондентского анализа как наиболее перспективный для использования при разработке метода построения классификаторов по обучающей информации, т.к. он позволяет использовать признаки, измеренные в наипростейшей шкале – шкале наименований, не налагая дополнительных требований на их типы данных и законы распределения, и получать при этом достаточно легко интерпретируемые представления взаимосвязей признаков с классами объектов в виде карт в пространстве небольшой размерности.

2. Для определения признаков, оказывающих значимое влияние на принадлежность объектов к классам, решено использовать методы множественных групповых сравнений, адекватные типу данных и шкале измерения каждого конкретного признака.

3. Проведенный анализ существующих способов формирования композиций классификаторов позволил определить эвристики, на основании комбинирования которых возможна разработка новых методов построения композиций. Наиболее перспективными в данной работе оказались комбинации методов взвешенного голосования, голосования по старшинству и смесей экспертов.

РАЗДЕЛ 3
РАЗРАБОТКА МЕТОДА ПОСТРОЕНИЯ КЛАССИФИКАТОРОВ НА ОСНОВАНИИ ГЕОМЕТРИЧЕСКОЙ ИНТЕРПРЕТАЦИИ СТРУКТУРЫ МНОГОМЕРНЫХ ДАННЫХ И МЕТОДОВ СОСТАВЛЕНИЯ КОМПОЗИЦИЙ КЛАССИФИКАТОРОВ

3.1 Разработка метода классификации на основе метрического подхода к геометрической интерпретации структуры данных

В данной работе использован метрический подход для решения задачи классификации с обучением, постановку которой в общем (упрощённом) виде можно представить следующим образом.

Есть конечное множество объектов $\{x_i\}_{i=1}^{L} \subset X$, каждый из которых известен по своему описанию с помощью некоторого набора признаков (переменных) $\{f_i\}_{i=1}^{n}$, которые можно представить как функции, действующие из пространства объектов в некоторое пространство значений: $f_i : X \to D_{f_i}$, где D_{f_i} – область значений признака f_i, X – называется пространством объектов, подмножество $\{x_i\}_{i=1}^{L} \subset X$ – обучающей выборкой.

Кроме того известно, что пространство объектов X некоторым образом разделено на классы $\{C_j\}_{j=1}^{m}$, и для каждого объекта из обучающей выборки $\{x_i\}_{i=1}^{L}$ известно, к какому классу он принадлежит.

Требуется построить классификатор, который будет относить

объекты к классам на основании значений их признаков $\{f_i\}_{i=1}^{n}$.

Как правило, при решении данной задачи классы представляются как некоторые подмножества (компактно расположенные области) пространства объектов [156–160]. В качестве координат объекта в пространстве объектов используются значения его признаков (или их преобразованных значений) [156–159, 161]. Классификаторы в этом случае строятся либо как разделяющие поверхности в этом пространстве [15, 162–164, 29], либо на основании оценок вероятностей принадлежности объекта к классам. Оценки, как правило, вычисляются на основе расстояния объекта в пространстве признаков до центра класса, ближайшего представителя в классе, и т.п. [158, 165–171]. Подобное представление показано на рис. 3.1.

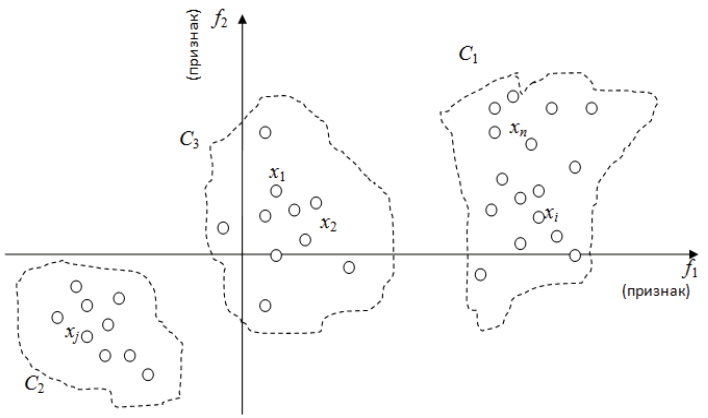


Рис. 3.1. Классическое представление о классах, объектах и признаках при решении задачи классификации

В отличие от упомянутого классического подхода, мы представляем класс, не как подмножество пространства объектов, а как ещё один из признаков, описывающих объект:

$c: X \to \{C_1, C_2, .. C_m\}$*.

В данном представлении строится математическая модель зависимости признака «класс» от остальных переменных. Разработка принципов построения математической модели происходила через последовательность получения ответов на ряд вопросов:

1. Насколько сильно влияет каждый из признаков на факт принадлежности объектов к классам? Т.е. каков вес (важность) каждого из признаков в классификации?
2. Все ли предикторы одинаково важны при прогнозировании разных классов? Насколько изменяется их вес (вклад) для различных классов?
3. Изменяется ли (и как изменяется) вес каждого предиктора в зависимости от его значений?

Таким образом, разработка метода классификации проходит этапы, схематично представленные на рис. 3.2.

Для ответа на 1-й вопрос, т.е. для оценки силы влияния предикторных показателей $\{f_i\}_{i=1}^n$ на выходной показатель c, используется подход, состоящий в выполнении следующих действий:

A. Приводим все признаки ($\{f_i\}_{i=1}^n$ и «класс» c) к единому масштабу. Это может быть достигнуто, например, нормировкой или стандартизацией количественных показателей, вычислением относительных частот категорий для качественных показателей.

---

*Подобное представление о классе использовалось в [172] при объяснении гипотез компактности и λ-компактности. При этом признак c назывался *целевым*, а $\{f_i\}_{i=1}^n$ – *описывающими* признаками.

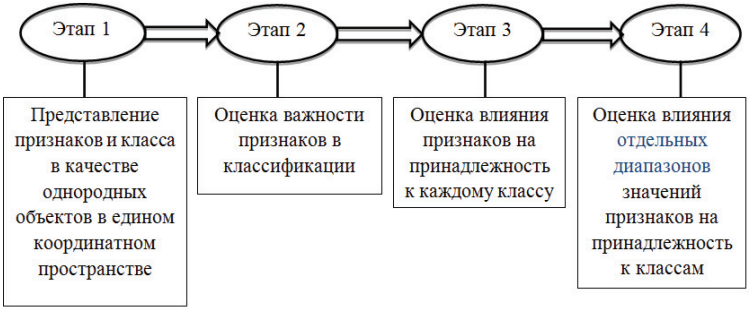


Рис. 3.2. Этапы разработки метода классификации

B. Исследуемые n+1 признак отображаются как точки в координатном пространстве размерности L, где каждая ось соответствует одному наблюдению из обучающей выборки объектов $\{x_i\}_{i=1}^{L}$. Данное представление показано на рис. 3.3.

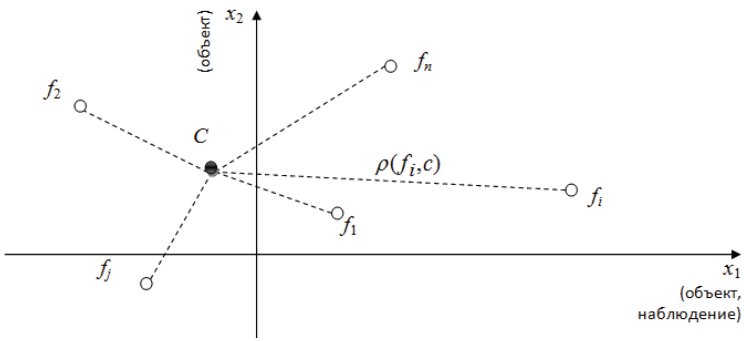


Рис. 3.3. Представление о классах, объектах и признаках при оценке важности признаков в классификации

C. Выбирается некоторая метрика ρ, с помощью которой вычисляются расстояния $\rho(f_i,c)$ от точек-предикторов f_i до точки-выходного показателя c.

D. Признакам, наиболее близким к c, назначаются бóльшие веса (важность) в классификации. По мере удаления точки f_i от точки c пропорционально увеличению расстояния уменьшается и вес признака f_i в математической модели определения переменной c («класс»).

Недостатки этого способа вычисления весовых коэффициентов признаков f_i для прогнозирования выходного показателя c обусловлены выбором осей координатного пространства и состоят в том, что:

– расстояния между точками, координатами которых являются наблюдения, не всегда будут характеризовать силу связи между переменными. Если оси пространства – это наблюдения, то близость переменных соответствует близости их значений, которая, не всегда говорит о сильной связи между показателями. Кроме того, достаточно сильные связи между переменными не всегда будут выражаться близостью их точек (например, при отрицательной корреляции).

– наблюдений, т.е. объясняющих измерений пространства, много, что приводит не только к усложнению вычислительной процедуры весовых коэффициентов, но и к более серьёзным последствиям, таким как переобучение метода [173, 75], (за счёт использования избыточной информации) и проявление действия т.н. «проклятия размерности» (чем больше слагаемых в сумме отклонений, характеризующей расстояние между точками, тем меньше различие между расстояниями) [174, 169].

Устранить указанные недостатки можно с помощью выбора

других координатных осей для пространства, в котором будут представлены исследуемые переменные. Для этого в качестве координат, объясняющих взаимное расположение показателей в общем пространстве, уместно использовать меры сходства (или различия) между их парами подобно тому, как это делается в методах анализа и упрощения геометрической структуры данных (многомерном шкалировании и корреспондентском анализе). При применении многомерного шкалирования (МШ) [175,176, 177], n+1 признаков первоначально отображаются точками в пространстве размерности $n \cdot (n+1)/2$ (каждая пара переменных даёт одно измерение), а затем с помощью монотонных преобразований размерность пространства сокращается таким образом, чтобы сходства/различия между точками, присутствующие в исходном пространстве, максимально сохранились. Процедура МШ требует на входе матрицу парных различий между переменными, не налагая при этом никаких ограничений на способ оценки этих различий. Мера сходства/различия между парой переменных может оцениваться на основании коэффициента корреляции, сопряжённости, совместной вероятности или даже быть результатом оценки эксперта в предметной области по его субъективному мнению.

Ценность подобного похода в том, что он позволяет измерять связи между переменными разной природы (в частности между качественными и количественными признаками) в одних терминах – в терминах расстояния между точками - их представителями в пространстве достаточно небольшой размерности [82].

В отличие от МШ многомерный корреспондентский анализ (КА, анализ соответствий) [97, 178], является методом упрощения структуры и геометрической интерпретации, разработанным

специально для данных, измеренных в номинальной шкале. Аналогом матрицы парных различий, используемой в МШ, здесь служит матрица, кросстабулирующая связи между всеми категориями всех исследуемых показателей (матрица Бёрта).

Таким образом, для более адекватной оценки силы влияния предикторных показателей $\{f_i\}_{i=1}^{n}$ на выходной показатель c, необходимо модифицировать алгоритм нахождения весовых коэффициентов предикторов следующим образом:

A. Формируется $(n+1)\times(n+1)$-матрица парных различий между всеми признаками ($\{f_i\}_{i=1}^{n}$ и «класс» c).

B. Используются методы упрощения геометрической структуры данных для отображения всех наших $n+1$ признаков в виде точек в координатном пространстве редуцированной размерности $r < n \cdot (n+1)/2$.

Пункты **C** и **D** сохраняем в первоначальной формулировке:

C. Выбирается некоторая метрика ρ, с помощью которой вычисляются расстояния $\rho(f_i, c)$ от точек-предикторов f_i до точки-выходного показателя c.

D. Весовые коэффициенты $\{\omega_i\}_{i=1}^{n}$ предикторов находятся как величины, обратно пропорциональные расстояниям точек f_i до точки c («класс»):

$$\omega_i = \left(\rho(f_i, c) \cdot \sum_{b=1}^{n} \frac{1}{\rho(f_b, c)} \right)^{-1} \qquad (3.1)$$

Формула (3.1) создана с учётом требования $\sum_{i=1}^{n} \omega_i = 1$, исходящего

из предположения о том, что набор предикторов $\{f_i\}_{i=1}^n$ является неизбыточным и не содержит зависимых переменных.

В предположении, что набор $\{f_i\}_{i=1}^n$ является предикторами выходного показателя «класс» (c), вносящими различный по величине вклад в определение этого выходного показателя, можем формально представить класс c как взвешенную сумму предикторов f_i:

$$\hat{c} = \sum_{i=1}^{n} \omega_i f_i + \delta, \qquad (3.2)$$

где δ – некоторое слагаемое, не обусловленное влиянием показателей f_i, включающее, в том числе, и случайную составляющую.

Аддитивность модели (3.2) предполагает отсутствие взаимодействий между предикторами f_i (т.е. переменные для использования в этой модели необходимо отбирать так, чтобы они были независимыми).

Предлагаемая модель (3.2) по форме записи похожа на модель регрессии, однако не является таковой, т.к. получаемая количественная оценка $\hat{c}$ не является ни приблизительной оценкой номера класса, ни оценкой вероятности принадлежности объекта к классу подобно модели логистической регрессии [179, 180]. Более корректным будет интерпретировать коэффициенты $\{\omega_i\}_{i=1}^n$ по аналогии с факторными нагрузками на показатели $\{f_i\}_{i=1}^n$, формирующими класс c как один общий фактор [32, 181–183]. Отличие предлагаемой модели от существующих состоит в способе нахождения коэффициентов ω_i, который основан на вычислении расстояний не между объектами (наблюдениями, центральными в классах, или ближайшими представителями классов), а между показателями, характеризующими

класс, и самим классом (как одним из признаков объекта), представленными в некотором обобщённом пространстве сходств/различий. Кроме того, при данном подходе за счёт предварительной подготовки данных и представления всех переменных в одном координатном пространстве реализуется возможность использования в одной модели как количественных, так и качественных предикторов.

Основными недостатками модели (3.2) (как и всех подобных регрессионных моделей) является линейное изменение выходного показателя в зависимости от предикторов и то, что предикторам придаётся одинаковый вес в прогнозировании различных классов.

Учёт нелинейности (и даже возможной немонотонности) изменения выходного показателя под влиянием предикторов в случае качественного выходного показателя «класс» достигается за счёт корректирования весовых коэффициентов предикторных переменных при прогнозировании разных классов.

Таким образом, для ответа на 2-й вопрос (о том, как изменяется вклад предикторов в прогнозирование разных классов) необходимо действовать так:

Вместо того чтобы рассматривать одну переменную «класс» (c) с m значениями $\{C_1, C_2, .. C_m\}$, рассмотрим m дихотомических переменных «класс_1» (c_1), «класс_2» (c_2), …, «класс_m» (c_m).

Способом, описанным выше, представим все показатели в виде точек в пространстве небольшой размерности (количество точек в таком представлении уже будет $n+m$, (рис. 3.4)) и получим своё уравнение регрессии для каждого выходного показателя c_j ($j = 1,\ldots,m$).

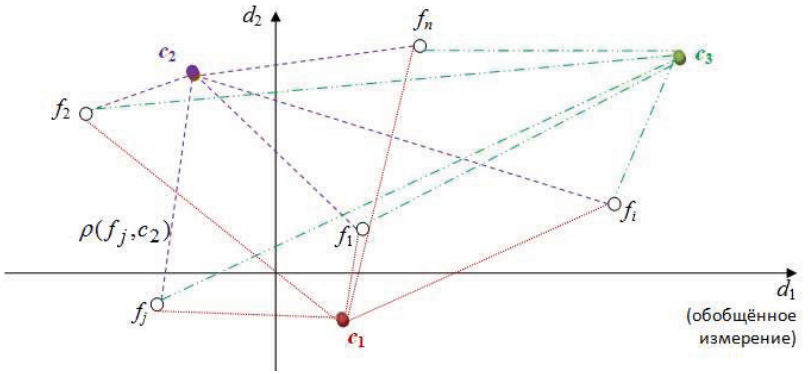


Рис. 3.4. Представление о классах, объектах и признаках при оценке влияния признаков на принадлежность объектов к различным классам

Таким образом, получаем систему из m уравнений для оценки принадлежности объекта к каждому из классов:

$$\hat{c}_j = \sum_{i=1}^{n} \omega_{ij} f_i + \delta_j \quad (\forall j = 1,\dots m) \tag{3.3}$$

Совокупность уравнений (3.3) по сути, является моделью алгоритма вычисления оценок (АВО [14, 119, 184, 118, 125]), на основании которых можно судить о принадлежности объекта к каждому из классов $\{C_j\}_{j=1}^{m}$. Подобный принцип построения моделей называется в литературе [18, 123] канонической моделью классификатора. Отличие же модели вычисления оценок (3.3) от существующих состоит снова в принципиально ином способе нахождения коэффициентов ω_{ij} для построения распознающего оператора.

Недостатком модели вычисления оценок на основании линейных комбинаций предикторов типа (3.3) является то, что в ней не учитывается возможная нелинейность и немонотонность

зависимости оценки c_j от изменения значений предикторов. Этот недостаток можно скорректировать, разбив область значений предикторной переменной на интервалы, на которых её поведение монотонно, и назначив этой переменной разные весовые коэффициенты на разных интервалах.

Таким образом, для ответа на вопрос 3 (о том, насколько изменяется влияние предикторов в зависимости от их поведения) необходимо поставить в соответствие каждому признаку f_i набор (категорий, интервалов) его возможных значений.

Если признаки представляются как функции на множестве объектов $x \in X : f_i : X \to D_{f_i}$, то после категоризации $f_i(x)$ будет задаваться как вектор

$$f_i(x) = \left(\mu_{g_{il}}(x)\right)_{l=1}^{k_i},$$

где $\mu_{g_{il}}(x)$ – характеристические функции подмножеств-категорий (интервалов) значений признака $g_{il} \subset D_{f_i}$, k_i – число подмножеств на которые разбивается область значений признака D_{f_i} $(1 < k_i \le m)$.

Следуя классическому определению характеристической функции [185–187]:

$$\mu_{g_{il}}(x) = \begin{cases} 1, & f_i(x) \in g_{il} \\ 0, & f_i(x) \notin g_{il} \end{cases}.$$

Однако, здесь возможно и использование других видов характеристических функций множества (например, нечётких [188–190]).

Далее в соответствии со способом, применяемым ранее, признаковое описание объектов представляется в виде конфигурации

из $N = m + \sum_{i=1}^{n} k_i \le m \cdot (n+1)$ точек пространства небольшой размерности, задающихся своими координатами. В полученной конфигурации отдельно позиционируются точки-представители классов c_j $(j = 1,\ldots, m)$, и отдельно – точки-представители (категорий) признаков υ_l $(l = 1,\ldots,\sum_{i=1}^{n} k_i)$, отвечающих за принадлежность объектов к классам.

Мера влияния определённой категории признака υ_l на принадлежность объекта к конкретному классу C_j оценивается как величина, обратная расстоянию точки-представителя этой категории до точки-представителя этого класса, нормированная на сумму расстояний всех точек-представителей категорий признаков:

$$\omega_{lj} = \left(\rho(v_l, c_j) \cdot \sum_{b=1}^{\sum_{i=1}^{n} k_i} \frac{1}{\rho(v_b, c_j)} \right)^{-1} \qquad (3.4)$$

Оценочные функции $W_j(x)$, характеризующие степень (вероятности) принадлежности некоторого объекта $x \in X$ к классам C_j $(j = 1,\ldots, m)$, вычисляются по формулам:

$$W_j(x) = \sum_{i=1}^{n} \sum_{l=1}^{k_i} \omega_{lj} \cdot \mu_{g_{li}}(x) \quad (\forall j = 1,\ldots,m). \qquad (3.5)$$

Наиболее вероятным классом для объекта x будет тот класс C_τ, для которого получено наибольшее значение функции $W_j(x)$:

$$\tau = \arg \max_{j=1,\ldots m} W_j(x). \qquad (3.6)$$

Таким образом, разработан метод и построена модель (3.4)—

(3.6) алгоритма классификации на основании вычисления оценок, которая позволяет учитывать влияние на принадлежность объектов к классам как качественных, так и количественных признаков, описывающих объект; а также, не смотря на представление формулы (3.5) в виде линейной комбинации предикторов, отражает нелинейность и немонотонность изменения оценки в зависимости от предикторных переменных. Отличие предлагаемой модели от существующих состоит в способе нахождения весовых коэффициентов переменных, объясняющих отнесение объектов к разным классам. Этот способ базируется на представлении класса не как подмножества пространства объектов, а как ещё одного показателя, составляющего признаковое описание объекта.

При решении практических задач построения классификаторов по обучающей информации и непосредственно классификации объектов, описанный метод классификации предлагается применять в соответствии со схемой, показанной на рис. 3.5. В данной схеме предполагается, что исходная обучающая информация (база данных) подаётся в виде матрицы признакового описания объектов. На первом этапе происходит структуризация и преобразование таблицы обучения, включающая и предварительную обработку данных. Матрица объектов-признаков сокращается путём исключения строк, отвечающих наблюдениям-выбросам, и столбцов, соответствующих коллинеарным (зависимым) переменным. Также исключаются столбцы матрицы, отвечающие показателям, не изменяющимся при переходе объектов из класса в класс. Такие показатели будем считать

не информативными для определения принадлежности объектов к классам в силу отсутствия влияния признака «класс» на их значения. Шаги по исключению коллинеарных и не информативных предикторов могут выполняться в любом порядке. Итогом выполнения этих процедур будет получение неизбыточного набора n независимых показателей, являющихся потенциальными предикторами, обуславливающими принадлежность объектов к разным классам.

Рис. 3.5. Алгоритм практического применения разработанного метода классификации

Категоризация значений предикторных переменных происходит путём разбиения областей значений количественных показателей на интервалы, характерные для различных классов, а для качественных

показателей – объединением (укрупнением) тех категорий их значений, которые характерны для одного класса. В результате для каждого класса будет сформирован набор т.н. эталонов, представляющих собой элементарные правила сравнения показателей с пороговыми значениями (для количественных признаков, описывающих объекты) либо перечень категорий, характерных для данного класса (для качественных признаков).

Итогом первого этапа является обучающая матрица «объект—признак», содержащая n+1 столбец, отвечающий классам объектов и категоризированным значениям n независимых предикторов. Данная матрица представляет собой входную информацию для получения геометрической интерпретации взаимосвязей между показателями с помощью методов корреспондентского анализа [97, 178].

Если обучающая информация задана в виде таблицы попарных сравнений, то этап предварительной обработки данных, включая приведение переменных к единой шкале, может быть опущен. В этом случае на этапе формирования пространственной структуры взаимосвязей признаков наиболее целесообразно использовать методы многомерного шкалирования [175, 176, 90, 89, 177].

На третьем этапе анализируется полученная пространственная структура и выбирается метрика, с помощью которой оцениваются расстояния между признаками на полученной карте. По формуле (3.4) вычисляются весовые коэффициенты категорий значений предикторных переменных в прогнозировании каждого класса. Тем самым при добавлении расчётных формул оценочных функций классов (3.5) и правила определения класса (3.6) формируется общая модель конкретного классификатора.

3.2 Разработка методов построения композиций классификаторов

На практике нередки ситуации, когда многочисленные попытки построения классификатора (ни увеличение числа признаков, описывающих объект, ни выбор алгоритма из различных семейств, ни применение к нему более эффективных методов обучения) не позволяют добиться приемлемого качества распознавания всех классов [127, 129]. Например, одно решающее правило даёт хорошие результаты при дискриминации объектов первого класса и неудовлетворительную точность определения принадлежности к другим классам; другие классификаторы – с большой точностью отделяют объекты других классов, имея низкую точность на объектах первого класса. Наиболее перспективным направлением в подобных ситуациях считается объединение нескольких алгоритмов в композицию (синонимы: комитет, ансамбль) с целью компенсирования их взаимных ошибок [138, 130, 134, 18], для чего стандартно используются три основных принципа голосования: простое, взвешенное и голосование по старшинству [149, 132, 131].

При простом голосовании строятся т.н. комитеты с логикой большинства, которые относят объект к тому классу, к которому его отнесли большинство, входящих в него классификаторов. Формально это может быть выражено соотношением [127, 145]:

$$T(x) = \frac{1}{m}\sum_{k=1}^{m} T^k(x),$$

где $\{T^k\}_{k=1}^{m}$ – набор базовых алгоритмов, входящих в комитет T.

При взвешенном голосовании, ответ каждого из классификаторов учитывается со своим весовым коэффициентом α_k, зависящим от качества данного алгоритма на обучающем множестве:

$$T(x) = \sum_{k=1}^{m} \alpha_k \cdot T^k(x).$$

В случае $\alpha_k = \ln(1 - p_k / p_k)$, $k = 1,\ldots m$, где p_k– ошибка классификатора T^k, получаем простейший линейный «наивный» байесовский классификатор [127, 191–193, 123]. В случае $\alpha_1 = \alpha_2 = \ldots \alpha_m = 1/m$ имеем дело с голосованием по большинству.

Развитием идеи взвешенного голосования являются т.н. смеси экспертов [153–155], в которых веса $\alpha_k = \alpha_k(x)$, $k = 1,..m$, т.е. являются не постоянными, а зависят от самогó классифицируемого объекта. При этом функции $\alpha_k(x)$ называются шлюзами или функциями компетентности [146, 152].

В комитетах с логикой старшинства [133, 149], называемых также машинами покрывающих множеств [150], происходит последовательная передача объекта от одного базового алгоритма к другому до тех пор, пока хотя бы один из них не выдаст ответ («покроет» объект). Этот метод предполагает последовательную одноклассовую классификацию. Т.е. первый алгоритм комитета отвечает за отнесение объекта к классу 1. Если он отказывается от классификации, то объект передаётся второму алгоритму, который может его отнести к классу 2. Если этого не произошло, объект передаётся к третьему классификатору и т.д. пока один из алгоритмов не примет решения.

3.2.1 Метод рейтингового голосования

В данном подпараграфе описана разработка алгоритма процедуры голосования смеси экспертов (взвешенного голосования) в

задаче классификации на несколько классов, основанной на вычислении рейтингов принадлежности классифицируемого образца к тому или иному классу. При расчёте рейтингов наравне с точностью алгоритмов смеси при прогнозировании отдельных классов учитываются и ошибки этих классификаторов на других классах.

Имеется конечное множество из m классификаторов $\{T^k\}_{k=1}^m$, решающих задачу дискриминации объектов $x \in X$ на n классов $\{C_i\}_{i=1}^n$. Для всех T^k известна (например, оценена по обучающей выборке) их точность прогнозирования классов, которая может быть задана квадратными $n \times n$-матрицами $P^k = \left(p_{ij}^k\right)_{i,j=1}^n$, где каждый элемент p_{ij}^k рассматривается как вероятность того, что объект, классифицированный алгоритмом T^k, как принадлежащий к классу C_j, в действительности принадлежит к классу C_i. Таким образом, диагональные элементы p_{ii}^k характеризуют точность алгоритмов T^k на классах C_i $(\forall k = 1,..m; \quad \forall i = 1,..n)$, а сумма элементов по каждому столбцу матриц P^k равна единице:

$$\sum_{i=1}^{n} p_{ij}^k = 1, \quad \forall j = 1,..n, \quad \forall k = 1,..m.$$

Результат действия множества построенных алгоритмов $\{T^k\}_{k=1}^m$ на некотором объекте x представляется бинарной индикаторной $n \times m$-матрицей

$$V(x) = \left(v_{ij}(x)\right)_{i=1,..n}^{j=1,..m}, \qquad \text{где } v_{ij}(x) = \begin{cases} 1, & \text{если} \quad T^j(x) = C_i \\ 0, & \text{если} \quad T^j(x) \neq C_i \end{cases} \qquad (3.7)$$

Матрица $V(x)$ обладает следующими свойствами:

i. Столбцы матрицы $V(x)$ соответствуют базовым алгоритмам, строки – прогнозируемым классам для объекта *x*.

ii. Суммы по строкам матрицы V представляют собой количество голосов в комитете, отданных за каждый класс:

$$\forall i = 1,..n: \quad 0 \le \sum_{j=1}^{m} v_{ij}(x) \le n \qquad (\forall x \in X).$$

iii. В каждом столбце матрицы V может быть не более одной единицы:

$$\forall j = 1,..m: \quad \sum_{i=1}^{n} v_{ij}(x) \le 1 \qquad (\forall x \in X).$$

Это утверждение соответствует тому, что каждый из базовых алгоритмов на одном объекте может давать не более одного ответа, т.е. отнести объект только к одному из классов или отказаться от его классификации. В первом случае: $\sum_{i=1}^{n} v_{ij}(x) = 1$. В случае отказа алгоритма T^j от классификации: $\sum_{i=1}^{n} v_{ij}(x) = 0$.

Далее формируем т.н. «рейтинговую» *n*×*m*-матрицу *W*, каждый столбец которой получается умножением матрицы P^k на *k*-й столбец матрицы *V*:

$$W^{\langle k \rangle}(x) = P^k \cdot V^{\langle k \rangle}(x) \tag{3.8}$$

Таким образом, элементы матрицы W $\forall i = 1,..n, \quad \forall k = 1,..m$ определяются соотношением:

$$w_{ik}(x) = \sum_{j=1}^{n} p_{ij}^{k} \cdot v_{jk}(x),$$

что, в силу свойств матрицы $V(x)$, эквивалентно:

$$w_{ik}(x) = \begin{cases} p_{i\theta}^{k}, & \text{если} \quad T^{k}(x) = C_{\theta}, \quad \text{где} \quad \theta \in \{1,..n\} \\ 0, & \text{если} \quad T^{k}(x) \text{ отказывается от классификации} \end{cases}.$$

Рейтинг принадлежности объекта x к классу C_i вычисляется как сумма по i-й строке матрицы W:

$$r_{C_i}(x) = \sum_{k=1}^{m} w_{ik}(x) = \sum_{k=1}^{m} p_{i\theta}^{k}, \qquad (3.9)$$

где $\theta \in \{0,1,\ldots n\}$– номер класса, к которому относит объект x алгоритм T^k (при этом под $\theta = 0$ подразумевается отказ алгоритма от классификации данного объекта).

Итоговым результатом процедуры рейтингового голосования комитета классификаторов T является класс, которому соответствует наибольший рейтинг:

$$T(x) = C_{\tau}, \qquad \text{где } \tau = \arg\max_{i=1,..n} r_{C_i}(x). \qquad (3.10)$$

Рассмотрим работу формализованной выше процедуры рейтингового голосования смеси алгоритмов классификации на примере.

Пусть для решения задачи классификации с обучением построено четыре классификатора $\{T^1, T^2, T^3, T^4\}$ для отнесения объектов к одному из трёх классов $\{C_1, C_2, C_3\}$. Для классификатора T^1 точность определения объектов класса C_1 равна 70%, при этом объекты первого класса могут быть ошибочно отнесены ко второму с вероятностью 5%, к третьему – с вероятностью 25%. Точность распознавания объектов второго класса классификатором T^1 равна 90%, при этом объекты класса C_2 могут быть ошибочно отнесены к классу C_1 с вероятностью 6%, к классу C_3 – с вероятностью 4%.

Объекты из класса C_3 алгоритм T^1 распознаёт с точностью 75%, вероятность отнести к третьему классу объект, в действительности принадлежащий к классу C_1, при этом равна 15%, а вероятность отнести к классу C_3 объект, в действительности принадлежащий к классу C_2, при этом равна 10%. Таким образом, для классификатора T^1 матрица точности прогнозирования:

$$P^1 = \begin{pmatrix} 0,7 & 0,06 & 0,15 \\ 0,05 & 0,9 & 0,1 \\ 0,25 & 0,04 & 0,75 \end{pmatrix}.$$

Классификатор T^2 определяет объекты первого класса с точностью 73%, второго – с точностью 85%, третьего – с точностью 90%. Ошибочно объекты первого класса могут быть отнесены к классу C_2 с вероятностью 10%, к классу C_3 с вероятностью 17%. Объекты, в действительности принадлежащие к классу C_2, могут быть отнесены этим классификатором к первому классу с вероятностью 6%, к классу C_3 – с вероятностью 9%. Объекты из класса C_3 с помощью алгоритма T^2 могут быть ошибочно классифицированы как лежащие в классе C_1 с вероятностью 8%, как лежащие в классе C_2 с вероятностью 2%. Т.е. для классификатора T^2 матрица точности прогнозирования:

$$P^2 = \begin{pmatrix} 0,73 & 0,06 & 0,08 \\ 0,1 & 0,85 & 0,02 \\ 0,17 & 0,09 & 0,9 \end{pmatrix}.$$

Аналогично, для классификаторов T^3 и T^4 примем такие матрицы прогнозирования:

$$P^3 = \begin{pmatrix} 0{,}83 & 0{,}02 & 0{,}08 \\ 0{,}06 & 0{,}95 & 0{,}25 \\ 0{,}11 & 0{,}03 & 0{,}67 \end{pmatrix},$$

$$P^4 = \begin{pmatrix} 0{,}63 & 0{,}03 & 0{,}09 \\ 0{,}32 & 0{,}91 & 0{,}2 \\ 0{,}05 & 0{,}06 & 0{,}71 \end{pmatrix}.$$

Пусть также на некотором классифицируемом объекте $\tilde{x}$ четыре построенных классификатора выдали следующие ответы:

$$\left\{T^1(\tilde{x}) = C_2; \quad T^2(\tilde{x}) = C_1; \quad T^3(\tilde{x}) = C_3; \quad T^4(\tilde{x}) = C_1\right\}.$$

Другими словами, первым алгоритмом объект был отнесён к классу C_2, классификаторы T^2 и T^4 оба относят объект к классу C_1, а алгоритм T^3 классифицирует этот же объект как принадлежащий к классу C_3. Т.е. матрица V (3.7) в этом случае имеет вид:

$$V(\tilde{x}) = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

В соответствии с простым голосованием по большинству объект $\tilde{x}$ следовало бы отнести к классу C_1. В предложенном же методе рейтингового голосования для получения окончательного ответа необходимо вычислить «рейтинговую» матрицу W по формуле (3.7). Так, первый столбец «рейтинговой» матрицы:

$$W^{\langle 1\rangle}(\tilde{x}) = P^1 \cdot V^{\langle 1\rangle}(\tilde{x}) = \begin{pmatrix} 0{,}7 & 0{,}06 & 0{,}15 \\ 0{,}05 & 0{,}9 & 0{,}1 \\ 0{,}25 & 0{,}04 & 0{,}75 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0{,}06 \\ 0{,}9 \\ 0{,}04 \end{pmatrix}.$$

Аналогично,

$$W^{\langle 2\rangle}(\widetilde{x}) = P^2 \cdot V^{\langle 2\rangle}(\widetilde{x}) = \begin{pmatrix} 0{,}73 & 0{,}06 & 0{,}08 \\ 0{,}1 & 0{,}85 & 0{,}02 \\ 0{,}17 & 0{,}09 & 0{,}9 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0{,}73 \\ 0{,}1 \\ 0{,}17 \end{pmatrix},$$

$$W^{\langle 3\rangle}(\widetilde{x}) = P^3 \cdot V^{\langle 3\rangle}(\widetilde{x}) = \begin{pmatrix} 0{,}83 & 0{,}02 & 0{,}08 \\ 0{,}06 & 0{,}95 & 0{,}25 \\ 0{,}11 & 0{,}03 & 0{,}67 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0{,}08 \\ 0{,}25 \\ 0{,}67 \end{pmatrix},$$

$$W^{\langle 4\rangle}(\widetilde{x}) = P^4 \cdot V^{\langle 4\rangle}(\widetilde{x}) = \begin{pmatrix} 0{,}63 & 0{,}03 & 0{,}09 \\ 0{,}32 & 0{,}91 & 0{,}2 \\ 0{,}05 & 0{,}06 & 0{,}71 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0{,}63 \\ 0{,}32 \\ 0{,}05 \end{pmatrix}.$$

Откуда $$W(\widetilde{x}) = \begin{pmatrix} 0{,}06 & 0{,}73 & 0{,}08 & 0{,}63 \\ 0{,}9 & 0{,}1 & 0{,}25 & 0{,}32 \\ 0{,}04 & 0{,}17 & 0{,}67 & 0{,}05 \end{pmatrix}.$$

Далее по формуле (3.9) находим рейтинги (оценки) принадлежности классифицируемого образца $\widetilde{x}$ к каждому из классов:

$$r_{C_1}(\widetilde{x}) = \sum_{k=1}^{4} w_{1k}(\widetilde{x}) = 0{,}06 + 0{,}73 + 0{,}08 + 0{,}63 = 1{,}5;$$

$$r_{C_2}(\widetilde{x}) = \sum_{k=1}^{4} w_{2k}(\widetilde{x}) = 0{,}9 + 0{,}1 + 0{,}25 + 0{,}32 = 1{,}57;$$

$$r_{C_3}(\widetilde{x}) = \sum_{k=1}^{4} w_{3k}(\widetilde{x}) = 0{,}04 + 0{,}17 + 0{,}67 + 0{,}05 = 0{,}93.$$

Как видим, $\max\limits_{i=1,2,3} r_{C_i}(\widetilde{x}) = 1{,}57 = r_{C_2}(\widetilde{x})$, следовательно, в соответствии с соотношением (3.10), объект $\widetilde{x}$ следует относить к классу C_2. Таким образом, полученный с помощью разработанного метода рейтингового голосования ответ отличается от ответа, полученного простым голосованием (по большинству), а также зависит от самого́ классифицируемого объекта, точнее, от матрицы

ответов базовых классификаторов, входящих в ансамбль, на этом объекте.

При программной реализации предложенного метода формирования композиций классификаторов математическая модель рейтингового голосования реализуется по алгоритму, представленному на рис. 3.6.

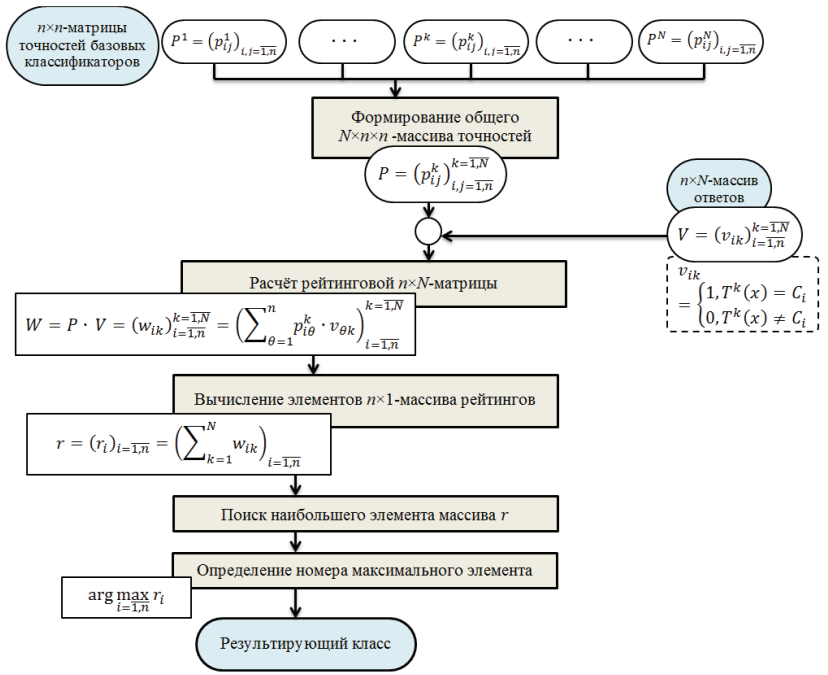


Рис. 3.6. Алгоритм программной реализации математической модели рейтингового голосования комитета классификаторов

На основании предварительно оцененных точностей N базовых классификаторов комитета в определении n классов формируется

единый $N \times n \times n$-массив P. С использованием полученного после работы всех базовых классификаторов бинарного массива ответов V вычисляются элементы $n \times N$ «рейтинговой матрицы» W. «Рейтинги» каждого из n классов вычисляются как суммы по строкам «рейтинговой матрицы». Далее из массива рейтингов выбирается максимальный элемент, и в качестве ответа выдаётся тот результирующий класс, которому он соответствует.

3.2.2 Рейтинговое голосование по старшинству

Настоящий подпараграф посвящён разработке алгоритма процедуры вынесения решения комитетом классификаторов, основанного на синтезе эвристик, лежащих в основе стандартных методов голосования по старшинству и взвешенного голосования.

Как и в комитете с логикой старшинства, предполагается, что число алгоритмов совпадает с количеством классов, и каждый классификатор наилучшим образом настроен на распознавание объектов именно «своего» класса. Т.е. любой базовый алгоритм T^i имеет максимальную точность распознавания объектов из класса C_i, что может быть отражено соотношением:

$$\forall i = 1,..n: \quad p_{ii}^i \geq p_{jj}^i \qquad \forall j \neq i, \quad j = 1,..n.$$

В стандартной процедуре голосования по старшинству сами базовые алгоритмы строятся как одноклассовые классификаторы. Каждый алгоритм обучается таким образом, чтобы максимально точно определять объекты только из одного (своего) класса [127]. Мы же рассматриваем более общую ситуацию, когда базовые алгоритмы обучались для определения в принципе всех классов, однако,

получается так, что их качество прогнозирования на одних классах лучше, чем на других. В этом случае последовательность $\{T^k\}_{k=1}^n$ ранжируется в порядке, соответствующем номерам классов, на которых они показали наибольшую точность распознавания. Т.е.

$$\forall i, j = 1,..n: \quad p_{ii}^i \geq p_{ii}^j, \tag{3.11}$$

причём, для $\forall i \neq j$ неравенство превращается в строгое (равенство достигается только в случае $i = j$).

В классической процедуре реализации машины покрывающих множеств [133, 149, 150], классифицируемый объект x подаётся сначала на вход первого базового алгоритма и, если $T^1(x) = C_1$, то объект x относится к классу C_1 и процедура на этом останавливается. Если же $T^1(x) \neq C_1$, то объект передаётся второму алгоритму, и т.д., пока не будет получен результат $T^i(x) = C_i$ и произойдёт остановка работы процедуры голосования комитета. Однако, остановка голосования при получении первого подходящего ответа, на наш взгляд, не всегда является оправданной, т.к. это увеличивает вероятность ошибки классификации, особенно для объектов, лежащих на границах классов, а также в случаях, когда характеристики качества распознавания различных классов одним и тем же базовым алгоритмом соизмеримы. Предлагаемая процедура голосования комитета классификаторов «рейтинговое голосование по старшинству» совмещает в себе принципы машины покрывающих множеств и смеси экспертов, позволяя повысить качество классификации в случаях, описанных выше.

Результат действия набора базовых классификаторов $\{T^k\}_{k=1}^n$ на

новом классифицируемом объекте x представляется матрицей $V(x)$ (3.7), обладающей свойствами (i)—(iii), которая, однако, в данном случае имеет размерность $n \times n$.

$$v_{ij}(x) = \begin{cases} 1, & \text{если} \quad T^j(x) = C_i \\ 0, & \text{если} \quad T^j(x) \neq C_i \end{cases}.$$

Т.к. каждый базовый классификатор T^k настроен таким образом, чтобы наилучший результат выдавать на классе C_k, то наиболее показательными характеристиками работы комитета в целом будут значения диагональных элементов матрицы V. Все принципиально различные ситуации расположения единиц на диагонали можно описать одним из трёх возможных значений следа матрицы $\mathbf{tr}V(x) = \sum_{i=1}^{n} v_{ii}(x)$.

Ситуация 1: $\mathbf{tr}V(x) = 0$. На диагонали матрицы V нет ни одной единицы, т.е. ни один из базовых классификаторов не идентифицировал образец как принадлежащий своему классу.

В этом случае целесообразным представляется отказаться от классификации данного объекта, как от нетипичного (попадающего в т.н. «область неуверенности», где все алгоритмы некомпетентны). Если подобных объектов обнаружится достаточно много, имеет смысл говорить о появлении (введении) нового класса, не рассматриваемого ранее [127, 15].

Результирующий ответ комитета: $T(x) = \varnothing$.

Ситуация 2: $\mathbf{tr}V(x) = 1$. На диагонали матрицы V только одна единица. Классическая ситуация для комитета с логикой старшинства, в которой он работает в чистом виде. В этом случае

ответом комитета будет тот класс C_τ, на котором эта единица появилась.

Результирующий ответ комитета: $T(x) = C_\tau$, где $\tau = \arg\{v_{ii} \quad | v_{ii}(x) = 1\}$.

Ситуация 3: $\mathbf{tr}V(x) > 1$. Несколько единиц появляются на диагонали матрицы V, когда сразу несколько базовых классификаторов высказываются за отнесение одного объекта x каждый к своему классу. В такой ситуации предлагается вычислить рейтинг каждого из спрогнозированных классов в соответствии с описанной выше процедурой рейтингового голосования и выдать результирующий ответ комитета: $T(x) = C_\tau$, где $\tau = \arg \max_i \sum_{j=1}^{\mathbf{tr}V(x)} p_{ji}^j$.

Вычисление рейтингов принадлежности объектов к классам осуществляется по алгоритму, описанному в предыдущем параграфе. Согласно соотношению (3.8) формируется «рейтинговая» $n \times n$ – матрица W, затем вычисляются суммы по строкам рейтинговой матрицы, представляющие собой оценки принадлежности классифицируемого образца к каждому из классов (формула (3.9)). Итоговым результатом процедуры рейтингового голосования комитета классификаторов T, согласно (3.10), является класс, которому соответствует наибольшая оценка (рейтинг).

Продемонстрируем, как работает описанная процедура рейтингового голосования по старшинству на следующем примере.

Пусть для решения задачи классификации с обучением построено четыре классификатора $\{T^1, T^2, T^3, T^4\}$ для отнесения объектов к одному из четырёх классов $\{C_1, C_2, C_3, C_4\}$. Для классификатора T^1 точность определения объектов класса C_1 равна 88%, при этом объекты первого класса могут быть ошибочно отнесены только к третьему четвёртому классам с вероятностями 3% и 9% соответственно. Точность распознавания объектов второго класса классификатором T^1 равна 60%, при этом объекты класса C_2 могут быть ошибочно отнесены к классу C_1 с вероятностью 8%, к классу C_3 – с вероятностью 21%, и к классу C_4 – с вероятностью 11%. Объекты из класса C_3 алгоритм T^1 распознаёт с точностью 63%. Вероятность отнести к третьему классу объект, в действительности принадлежащий к классу C_1, при этом равна 4%; вероятность отнести к классу C_3 объект, в действительности принадлежащий к классу C_2, при этом равна 17%, а вероятность отнести к классу C_3 объект из класса C_4 составляет16%. Для объекта из класса C_4 вероятность того, что он будет правильно классифицирован алгоритмом T^1, равна 69%. Ошибки классификации этого алгоритма для объектов четвёртого класса таковы: к классу C_1 объект может быть ошибочно отнесён с вероятностью 10%, к классу C_2 – с вероятностью 12%, к классу C_3 – с вероятностью 9%. Таким образом, для классификатора T^1 матрица точности прогнозирования:

$$P^1 = \begin{pmatrix} 0{,}88 & 0{,}08 & 0{,}04 & 0{,}1 \\ 0 & 0{,}6 & 0{,}17 & 0{,}12 \\ 0{,}03 & 0{,}21 & 0{,}63 & 0{,}09 \\ 0{,}09 & 0{,}11 & 0{,}16 & 0{,}69 \end{pmatrix}.$$

Аналогично определяются матрицы точности классификации для алгоритмов T^2, T^3 и T^4:

$$P^2 = \begin{pmatrix} 0{,}7 & 0{,}01 & 0{,}12 & 0{,}07 \\ 0{,}05 & 0{,}85 & 0{,}05 & 0{,}09 \\ 0{,}15 & 0{,}11 & 0{,}65 & 0{,}18 \\ 0{,}1 & 0{,}03 & 0{,}18 & 0{,}66 \end{pmatrix},$$

$$P^3 = \begin{pmatrix} 0{,}75 & 0{,}15 & 0{,}01 & 0{,}05 \\ 0{,}04 & 0{,}71 & 0{,}08 & 0{,}1 \\ 0{,}19 & 0{,}14 & 0{,}8 & 0{,}12 \\ 0{,}02 & 0 & 0{,}11 & 0{,}73 \end{pmatrix},$$

$$P^4 = \begin{pmatrix} 0{,}64 & 0{,}01 & 0{,}2 & 0{,}08 \\ 0{,}12 & 0{,}61 & 0{,}07 & 0{,}04 \\ 0{,}21 & 0{,}11 & 0{,}72 & 0{,}11 \\ 0{,}03 & 0{,}27 & 0{,}01 & 0{,}77 \end{pmatrix}.$$

Обратим внимание, что базовые алгоритмы классификации предварительно ранжированы таким образом, что алгоритм T^i имеет максимальную точность на классе C_i (на «своём» классе) $\forall i = 1,2,3,4$, т.е. выполняется свойство (3.11). Более того, они расположены в порядке убывания точности на своих классах, т.е.

$$\forall i, j = 1,..4: \quad i < j \Leftrightarrow p_{ii}^i > p_{jj}^j .$$

Пусть также на некотором классифицируемом объекте $\tilde{x}$ четыре построенных классификатора сработали так, что каждый из них отнёс объект к своему классу:

$$\forall i, j = 1,..4: \quad T^i(\tilde{x}) = C_i .$$

Матрица ответов V (3.7) в этом случае принимает вид единичной матрицы 4×4:

$$V(\tilde{x}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

В данном случае, если бы применялся стандартный алгоритм голосования по старшинству, процесс работы комитета остановился сразу же после работы первого классификатора T^1. В этом случае классифицируемый образец следовало бы отнести к первому классу (C_1).

В случае же модифицированной нами процедуры, названной рейтинговым голосованием по старшинству, необходимо прогнозировать, что $\tilde{x} \in C_3$.

Действительно, вычислим элементы столбцов и составим рейтинговую матрицу *W*, применив формулу (3.8):

$$W^{\langle 1 \rangle}(\tilde{x}) = P^1 \cdot V^{\langle 1 \rangle}(\tilde{x}) = \begin{pmatrix} 0{,}88 & 0{,}08 & 0{,}04 & 0{,}1 \\ 0 & 0{,}6 & 0{,}17 & 0{,}12 \\ 0{,}03 & 0{,}21 & 0{,}63 & 0{,}09 \\ 0{,}09 & 0{,}11 & 0{,}16 & 0{,}69 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0{,}88 \\ 0 \\ 0{,}03 \\ 0{,}09 \end{pmatrix},$$

$$W^{\langle 2 \rangle}(\tilde{x}) = P^2 \cdot V^{\langle 2 \rangle}(\tilde{x}) = \begin{pmatrix} 0{,}7 & 0{,}01 & 0{,}12 & 0{,}07 \\ 0{,}05 & 0{,}85 & 0{,}05 & 0{,}09 \\ 0{,}15 & 0{,}11 & 0{,}65 & 0{,}18 \\ 0{,}1 & 0{,}03 & 0{,}18 & 0{,}66 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0{,}01 \\ 0{,}85 \\ 0{,}11 \\ 0{,}03 \end{pmatrix},$$

$$W^{\langle 3 \rangle}(\tilde{x}) = P^3 \cdot V^{\langle 3 \rangle}(\tilde{x}) = \begin{pmatrix} 0{,}75 & 0{,}15 & 0{,}01 & 0{,}05 \\ 0{,}04 & 0{,}71 & 0{,}08 & 0{,}1 \\ 0{,}19 & 0{,}14 & 0{,}8 & 0{,}12 \\ 0{,}02 & 0 & 0{,}11 & 0{,}73 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0{,}01 \\ 0{,}08 \\ 0{,}8 \\ 0{,}11 \end{pmatrix},$$

$$W^{\langle 4\rangle}(\tilde{x}) = P^4 \cdot V^{\langle 4\rangle}(\tilde{x}) = \begin{pmatrix} 0{,}64 & 0{,}01 & 0{,}2 & 0{,}08 \\ 0{,}12 & 0{,}61 & 0{,}07 & 0{,}04 \\ 0{,}21 & 0{,}11 & 0{,}72 & 0{,}11 \\ 0{,}03 & 0{,}27 & 0{,}01 & 0{,}77 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0{,}08 \\ 0{,}04 \\ 0{,}11 \\ 0{,}77 \end{pmatrix}.$$

Откуда $$W(\tilde{x}) = \begin{pmatrix} 0{,}88 & 0{,}01 & 0{,}01 & 0{,}08 \\ 0 & 0{,}85 & 0{,}08 & 0{,}04 \\ 0{,}03 & 0{,}11 & 0{,}8 & 0{,}11 \\ 0{,}09 & 0{,}03 & 0{,}11 & 0{,}77 \end{pmatrix}.$$

Далее по формуле (3.9) находим рейтинги (оценки) принадлежности классифицируемого образца $\tilde{x}$ к каждому из классов:

$$r_{C_1}(\tilde{x}) = \sum_{k=1}^{4} w_{1k}(\tilde{x}) = 0{,}88 + 0{,}01 + 0{,}01 + 0{,}08 = 0{,}98,$$

$$r_{C_2}(\tilde{x}) = \sum_{k=1}^{4} w_{2k}(\tilde{x}) = 0 + 0{,}85 + 0{,}08 + 0{,}04 = 0{,}97,$$

$$r_{C_3}(\tilde{x}) = \sum_{k=1}^{4} w_{3k}(\tilde{x}) = 0{,}03 + 0{,}11 + 0{,}8 + 0{,}11 = 1{,}05,$$

$$r_{C_4}(\tilde{x}) = \sum_{k=1}^{4} w_{4k}(\tilde{x}) = 0{,}09 + 0{,}03 + 0{,}11 + 0{,}77 = 1.$$

Как видим, $\max\limits_{i=1,..4} r_{C_i}(\tilde{x}) = 1{,}05 = r_{C_3}(\tilde{x})$, следовательно, в соответствии с соотношением (3.10), объект $\tilde{x}$ следует относить к классу C_3.

Рассмотрим другую ситуацию, когда на некотором образце $\hat{x} \in X$ базовые алгоритмы комитета сработали таким образом, что классификаторы T^2 и T^3 отнесли его к третьему классу, а классификаторы T^1 и T^4 – к четвёртому:

$$T^2(\hat{x}) = T^3(\hat{x}) = C_3; \qquad T^1(\hat{x}) = T^4(\hat{x}) = C_4.$$

Т.е. матрица ответов выглядит следующим образом:

$$V(\hat{x}) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

В данном случае, при стандартном голосовании по старшинству классифицируемый образец следовало бы отнести к классу C_3.

Применение голосования по большинству не дало бы удовлетворительного ответа. Т.к. для двух классов получено одинаковое количество ответов, простой комитет с логикой большинства отказался бы от классификации.

Применим рейтинговое голосование по старшинству, в соответствии с которым вычислим сначала рейтинговую матрицу W:

$$W^{\langle 1\rangle}(\hat{x}) = P^1 \cdot V^{\langle 1\rangle}(\hat{x}) = \begin{pmatrix} 0{,}88 & 0{,}08 & 0{,}04 & 0{,}1 \\ 0 & 0{,}6 & 0{,}17 & 0{,}12 \\ 0{,}03 & 0{,}21 & 0{,}63 & 0{,}09 \\ 0{,}09 & 0{,}11 & 0{,}16 & 0{,}69 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0{,}1 \\ 0{,}12 \\ 0{,}09 \\ 0{,}69 \end{pmatrix},$$

$$W^{\langle 2\rangle}(\hat{x}) = P^2 \cdot V^{\langle 2\rangle}(\hat{x}) = \begin{pmatrix} 0{,}7 & 0{,}01 & 0{,}12 & 0{,}07 \\ 0{,}05 & 0{,}85 & 0{,}05 & 0{,}09 \\ 0{,}15 & 0{,}11 & 0{,}65 & 0{,}18 \\ 0{,}1 & 0{,}03 & 0{,}18 & 0{,}66 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0{,}12 \\ 0{,}05 \\ 0{,}65 \\ 0{,}18 \end{pmatrix},$$

$$W^{\langle 3\rangle}(\hat{x}) = P^3 \cdot V^{\langle 3\rangle}(\hat{x}) = \begin{pmatrix} 0{,}75 & 0{,}15 & 0{,}01 & 0{,}05 \\ 0{,}04 & 0{,}71 & 0{,}08 & 0{,}1 \\ 0{,}19 & 0{,}14 & 0{,}8 & 0{,}12 \\ 0{,}02 & 0 & 0{,}11 & 0{,}73 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0{,}01 \\ 0{,}08 \\ 0{,}8 \\ 0{,}11 \end{pmatrix},$$

$$W^{\langle 4\rangle}(\hat{x}) = P^4 \cdot V^{\langle 4\rangle}(\hat{x}) = \begin{pmatrix} 0{,}64 & 0{,}01 & 0{,}2 & 0{,}08 \\ 0{,}12 & 0{,}61 & 0{,}07 & 0{,}04 \\ 0{,}21 & 0{,}11 & 0{,}72 & 0{,}11 \\ 0{,}03 & 0{,}27 & 0{,}01 & 0{,}77 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0{,}08 \\ 0{,}04 \\ 0{,}11 \\ 0{,}77 \end{pmatrix}.$$

Следовательно,

$$W(\hat{x}) = \begin{pmatrix} 0,1 & 0,12 & 0,01 & 0,08 \\ 0,12 & 0,05 & 0,08 & 0,04 \\ 0,09 & 0,65 & 0,8 & 0,11 \\ 0,69 & 0,18 & 0,11 & 0,77 \end{pmatrix}.$$

Найдём оценки принадлежности объекта $\hat{x}$ к классам по формуле (3.9):

$$r_{C_1}(\hat{x}) = \sum_{k=1}^{4} w_{1k}(\tilde{x}) = 0,1 + 0,12 + 0,01 + 0,08 = 0,31,$$

$$r_{C_2}(\hat{x}) = \sum_{k=1}^{4} w_{2k}(\tilde{x}) = 0,12 + 0,05 + 0,08 + 0,04 = 0,29,$$

$$r_{C_3}(\hat{x}) = \sum_{k=1}^{4} w_{3k}(\tilde{x}) = 0,09 + 0,65 + 0,8 + 0,11 = 1,65,$$

$$r_{C_4}(\hat{x}) = \sum_{k=1}^{4} w_{4k}(\tilde{x}) = 0,69 + 0,18 + 0,11 + 0,77 = 1,75.$$

Как видим, $\max\limits_{i=1,..4} r_{C_i}(\hat{x}) = 1,75 = r_{C_4}(\hat{x})$, следовательно, в соответствии с (3.10), необходимо прогнозировать, что $\hat{x} \in C_4$.

Как показывают приведенные примеры, ответы, полученные с помощью формализованной в данном параграфе процедуры голосования комитета классификаторов «рейтинговое голосование по старшинству», отличаются от ответов, выдаваемых стандартной машиной покрывающих множеств и комитетами с логикой большинства. Более того, применение рейтингового голосования по старшинству позволяет получать ответы (избегать отказа от классификации) в случаях получения от базовых алгоритмов равного количества голосов за несколько классов на одном объекте.

Алгоритм выполнения процедуры рейтингового голосования по старшинству реализуется в соответствии со схемой на рис. 3.7.

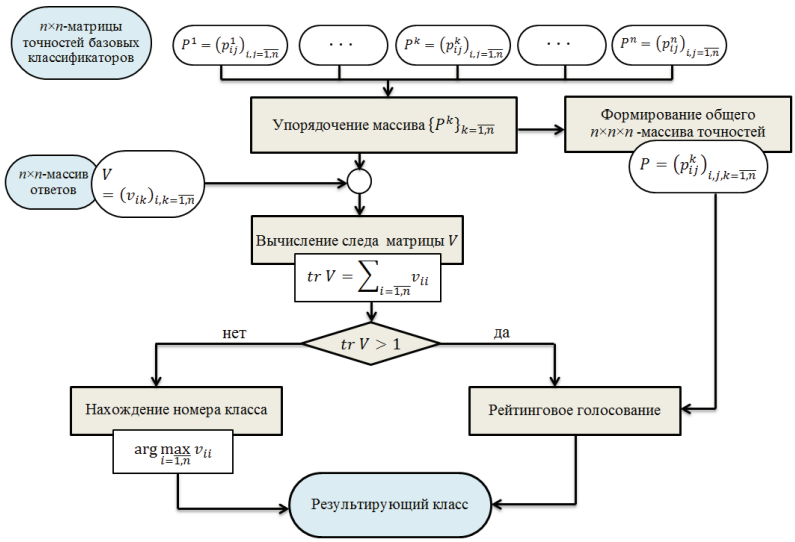


Рис. 3.7. Алгоритм программной реализации математической модели рейтингового голосования по старшинству

Здесь количество базовых классификаторов ансамбля совпадает с количеством распознаваемых классов ($N = n$). На начальном этапе происходит упорядочение набора базовых классификаторов в соответствии с их точностями распознавания различных классов. Перенумерация базовых классификаторов и классов выполняется таким образом, чтобы для элементов общего $n \times n \times n$-массива точностей P выполнялись свойства:

1) $\forall i, j = 1,\ldots,n: \quad p_{ii}^i > p_{jj}^i \quad \forall j \neq i$. Т.е. устанавливается соответствие между номерами классификаторов и номерами классов. Тем самым каждому классу ставится в соответствие «свой» классификатор, который распознаёт объекты из этого класса лучше других.

2) $\forall i, j = 1,\ldots,n:\quad i < j \;\Leftrightarrow\; p_{ii}^{i} > p_{jj}^{j}$, что соответствует расположению базовых классификаторов в порядке убывания точностей определения «своих» классов.

По имеющемуся $n \times n$-массиву ответов V вычисляется сумма его диагональных элементов (след матрицы), отражающая количество классификаторов, проголосовавших за «свои» классы. Если таких классификаторов более одного, выполняется процедура рейтингового голосования в соответствии со схемой на рис. 3.6. В противном случае определяется номер классификатора, который сработал на «своём» классе, и этот класс выдаётся пользователю в качестве окончательного ответа.

3.3 Разработка информационной технологии для задач классификации в медицинских приложениях

Разработанные в данном разделе метод построения классификаторов на основании геометрической интерпретации структуры многомерных данных и методы составления композиций классификаторов предложено использовать в качестве математического обеспечения в информационных технологиях поддержки принятия решений при дифференциальной диагностике различных заболеваний, оценки степени тяжести состояния пациентов, прогнозировании летальности и других прикладных медицинских задачах, позволяющих постановку в виде задачи классификации с обучением. Предложенные методы могут использоваться в рамках одной информационной технологии, либо

отдельно для построения математических моделей классификаторов или формирования алгоритмических композиций.

Технология классификации, в которой используется только метод построения классификатора, схематично представлена на рис. 3.8. Все этапы информационной технологии реализуются соответствующим программным обеспечением, разработанным для предварительно построенной математической модели классификатора.

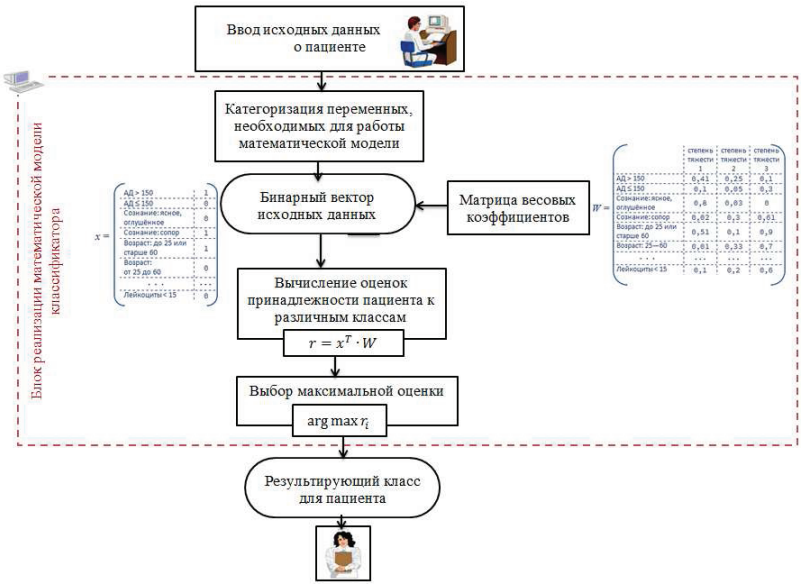


Рис. 3.8. Информационная технология на основе одного классификатора

На начальном этапе (рис. 3.8) вводятся данные о конкретном пациенте, для которого необходимо спрогнозировать принадлежность к одному из классов. Значения необходимых для работы

математической модели показателей преобразуются в бинарный вектор (x) в соответствии с элементарными правилами (шаблонами), используемыми классификатором. Сформированный бинарный вектор исходных данных умножается на матрицу весовых коэффициентов шаблонов (W), вычисленную на основании графической интерпретации взаимного расположения классов и объясняющих элементарных правил. В результате умножения получается вектор оценок (r), элементы которого отражают, насколько принадлежность к каждому из классов характерна (вероятна) для данного пациента. Результирующий прогноз класса, выдаваемый пользователю, получается путём выбора максимального элемента оценочного вектора r.

Информационная технология классификации для случая нескольких моделей, объединённых в ансамбль, представлена на рис. 3.9. После ввода исходных данных о пациенте реализуются математические модели всех базовых классификаторов в соответствии со схемой 3.8 и получается набор прогнозов принадлежности к классам для данного пациента. Ответы, полученные в результате применения различных классификаторов, могут, как совпадать, так и быть разными. Для получения результирующего класса имеющийся набор ответов записывается в виде бинарного массива, который затем используется в математической модели составления комитета классификаторов. Столбцы массива ответов соответствуют базовым классификаторам, строки – возможным классам для пациента. На пересечении каждых строки и столбца стоит 1, если соответствующий базовый классификатор отнёс пациента к соответствующему классу; 0 – в противном случае.

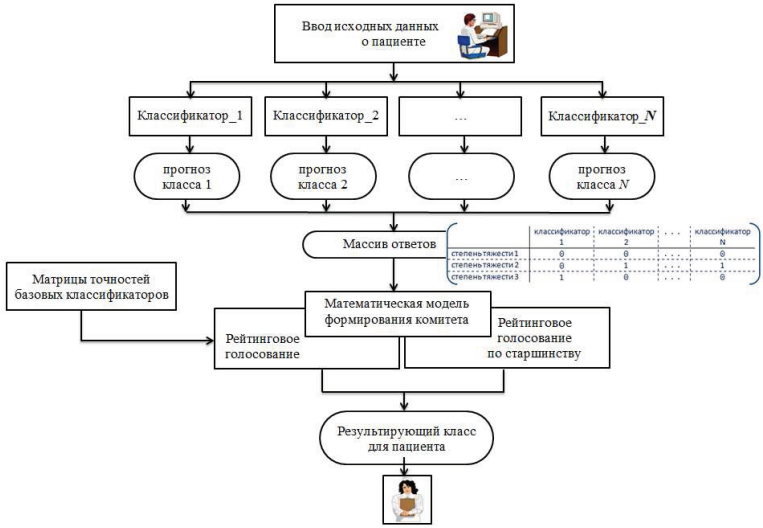


Рис. 3.9. Информационная технология на основе ансамбля классификаторов

В описываемой информационной технологии возможно применение двух различных математических моделей формирования комитета, первая из которых составлена на основе рейтингового голосования, вторая – рейтингового голосования по старшинству. Независимо от того, какая математическая модель реализуется, для её работы необходимо использование матриц прогностических точностей и ошибок базовых классификаторов при определении различных степеней тяжести состояния. Результатом работы математической модели голосования комитета классификаторов является окончательный прогноз класса, к которому следует отнести пациента, который и выдаётся пользователю (врачу).

Разработка представляемой информационной технологии включает 6 этапов, представленных на рис. 3.10. На вход первого

этапа подаётся обучающая информация, которая в данной схеме, как и в большинстве практических приложений, представляется в виде признакового описания конечного набора объектов (пациентов). Предполагается существование нескольких непересекающихся классов объектов (диагнозов, исходов лечения, степеней тяжести состояния пациентов, или т.п.). В исходной обучающей выборке известно разделение объектов (пациентов) на классы. Перед непосредственно применением разработанного метода построения классификаторов, обучающая информация проходит этап предварительной обработки. Основной целью этого этапа является извлечение из имеющихся данных информативных элементарных правил, т.е. выбора шаблонов (паттернов, простейших множеств) [194—197, 160], характеризующих причины принадлежности объектов к конкретным классам. Шаблоны, как правило, представляются для признаков, измеренных как минимум в интервальной шкале, в виде сравнений с некоторыми пороговыми значениями; для качественных признаков – в виде перечня категорий, характерных для определённых классов. Полученные элементарные правила рассматриваются как категоризированные значения признаков, что позволяет интерпретировать все объясняющие переменные как качественные, каждую со своим набором категорий.

Ключевым моментом в разработке информационной технологии является работа с классом не как с подмножеством обучающей выборки, а как с ещё одной качественной переменной, составляющей признаковое описание объекта. Такой подход даёт возможность рассматривать и объясняющие переменные, и классы, как однородные показатели (измеренные в одинаковой шкале).

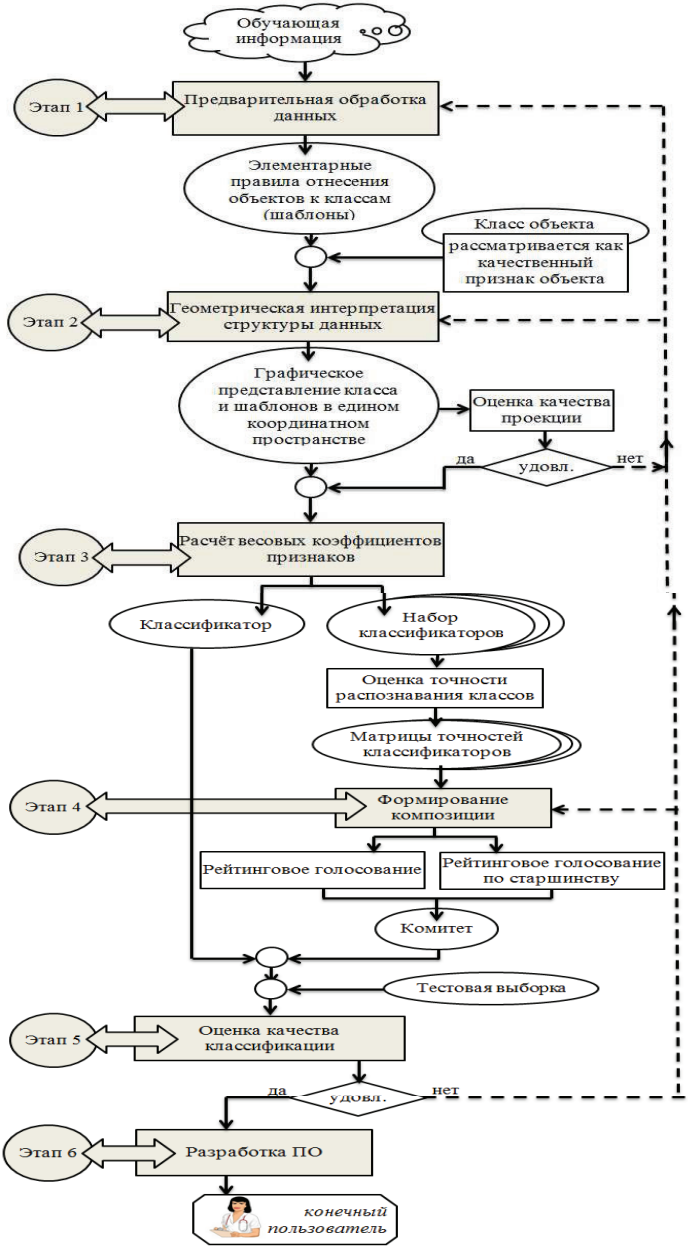


Рис. 3.10. Этапы разработки информационной технологии классификации

На втором этапе с помощью методов геометрической интерпретации и упрощения структуры данных, [76–78, 15, 177], получаются карты представления признаков и классов в (некотором обобщённом) едином координатном пространстве. Для описанного выше представления обучающей информации и результатов её предварительной обработки наиболее целесообразным представляется применение в качестве метода геометрической интерпретации множественного анализа соответствий (корреспондентского анализа) [15, 97, 178].

Полученное пространственное представление обязательно подвергается проверке на точность сохранения связей между отображаемыми показателями в обобщённом пространстве сокращённой размерности. При неудовлетворительном качестве полученной проекции необходимо вернуться к предыдущим этапам и пересмотреть условия применения метода (например, изменить количество измерений пространства окончательной конфигурации), или переформулировать элементарные правила (пересмотреть, уточнить пороговые значения, правила укрупнения и разбиения категорий, и т.п.). Если эти подходы всё равно не позволят добиться приемлемого качества проекции, то возможно получение нескольких независимых карт (например, для каждого класса в отдельности).

На третьем этапе рассчитываются вклады каждого из шаблонов в принадлежность объекта к определённому классу. Правила расчёта весовых коэффициентов, отражающих эти вклады, и их разработка детально описаны в параграфе 3.1 настоящей работы.

Таким образом, на выходе этапа 3 получается либо математическая модель одного классификатора, либо их набор. В случае нескольких моделей, они затем объединяются в ансамбль с

помощью одного из разработанных методов построения композиций классификаторов. В работе предложено и формализовано два метода построения композиций: т.н. «рейтинговое голосование», являющееся одной из возможных реализаций смеси экспертов; и «рейтинговое голосование по старшинству». Последнее модифицирует метод «рейтингового голосования», добавляя в него принципы комитетов с логикой старшинства, что в некоторых практических приложениях является более эффективным. Подробно процедуры «рейтингового голосования» и «рейтингового голосования по старшинству» описаны в параграфах 3.2.1 и 3.2.2.

Используемые при разработке информационной технологии методы формирования ансамбля (комитета) классификаторов в качестве входной информации требуют наличия для каждого базового классификатора матрицы оценок его точности распознавания и ошибок на объектах различных классов. Указанные матрицы могут быть составлены как по обучающей выборке, так и с использованием дополнительной тестовой выборки. Обратим также внимание на то, что под термином «точность распознавания классов», в отличие от употребляемого более часто термина *точность производителя* (producer's accuracy), в данной технологии понимается т.н. *точность пользователя* (user's accuracy) [28]. Если producer's accuracy определяется как отношение правильно распознанных классификатором образцов из определённого класса к общему числу объектов в этом классе, то user's accuracy вычисляется как отношение правильно распознанных образцов класса к общему количеству образцов, отнесённых классификатором к этому классу. Таким образом, как нам кажется, использование в данной информационной технологии точности пользователя для оценки процента правильных и

ошибочных ответов базовых классификаторов на различных классах позволяет более корректно оценить качество их будущей совместной работы.

Итак, на выходе четвёртого этапа технологии имеется комитет алгоритмов классификации, сформированный одним из способов: с помощью «рейтингового голосования» или «рейтингового голосования по старшинству». Очевидно, что в случае получения на выходе третьего этапа единственного классификатора, этап 4 данной информационной технологии не выполняется.

На пятом этапе по тестовой выборке происходит оценка качества распознавания классов, предоставляемого разработанной математической моделью классификатора или комитета классификаторов.

При неудовлетворительном качестве классификации, в случае комитета классификаторов, возможно возвращение к четвёртому этапу и применение других методов составления алгоритмических композиций. Во всех случаях добиться приемлемого качества классификации возможно возвращением на этапы 2 или 1, которое позволит получить новые графические представления либо формулировки элементарных шаблонов отнесения объектов к классам.

Если же точность классификации, оцененная на пятом этапе по тестовой выборке, признана приемлемой, то реализуется шестой (завершающий) этап разработки информационной технологии. Этот этап состоит в создании программного обеспечения, реализующего математическую модель классификатора и алгоритм голосования ансамбля. Разработанное в рамках предлагаемой информационной

технологии программное обеспечение предоставляется конечному пользователю для практического применения.

Таким образом, в качестве основных результатов данного раздела можно выделить следующие положения.

1. Разработанный метод классификации, использующий представление о классе, не как о подмножестве пространства объектов, а как об одной из составляющих признакового описания объекта, а также способ нахождения весовых коэффициентов предикторов на основании метрического подхода и геометрической интерпретации структуры данных, позволяет построение моделей классификаторов, в которых возможно совместное использование как количественных, так и качественных объясняющих переменных с учётом нелинейности и немонотонности их поведения при переходе от класса к классу.

2. Разработанный метод составления ансамблей классификаторов «рейтинговое голосование», являющийся одной из реализаций метода смесей экспертов, позволяет повысить качество классификации в случаях совместного использования нескольких алгоритмов, что достигается за счёт совместного учёта не только точности базовых алгоритмов при прогнозировании отдельных классов, а также и их ошибок на других классах.

3. Предложенный алгоритм вынесения решения комитетом классификаторов «рейтинговое голосование по старшинству», основанный на совмещении принципов взвешенного голосования и

машин покрывающих множеств, применение которого целесообразно в случаях, когда базовые классификаторы можно ранжировать в порядке их точности на различных классах, даёт возможность усовершенствовать стандартную процедуру голосования по старшинству, а также уменьшить количество отказов от классификации по сравнению со стандартными комитетами с логикой старшинства или большинства.

4. Предложенный алгоритм совместного использования представленных в работе методов построения классификаторов и формирования их ансамблей позволяет построение математических моделей, на основе которых реализуются составляющие информационной технологии реализации моделей классификации и их композиций для медицинских приложений.

РАЗДЕЛ 4
РЕЗУЛЬТАТЫ ПРАКТИЧЕСКОГО ПРИМЕНЕНИЯ РАЗРАБОТАННЫХ МЕТОДОВ ПОСТРОЕНИЯ КЛАССИФИКАТОРОВ И ИХ КОМПОЗИЦИЙ

4.1 Оценка тяжести состояния пациентов и прогноз исхода при травматических повреждениях поджелудочной железы и травматическом панкреатите

4.1.1 Характеристика входных данных

Для построения математических моделей использовались показатели пациентов с травмами ПЖ и травматическим панкреатитом, проходивших лечение в четырёх ургентных клиниках г. Харькова (Украина) за период 2007—2012 гг. Все пациенты были классифицированы экспертом (практикующим хирургом, специалистом в области хирургического лечения поджелудочной железы) в зависимости от тяжести их состояния. При такой апостериорной экспертной оценке тяжести состояния пациента учитывались следующие факторы: тяжесть повреждений, исход, наличие послеоперационных осложнений, также, наличие сопутствующих заболеваний, кровопотеря (уровень гемоглобина) и др. Тяжесть повреждений оценивалась в соответствии с международной шкалой оценки тяжести травмы AIS (Abbreviated Injury Scale) применимо к травме живота, но полученный результат не всегда совпадал с экспертной оценкой тяжести состояния пациента, так как эксперт учитывал и дополнительные факторы. В результате в зависимости от тяжести состояния и тяжести повреждений были сформированы следующие группы пациентов:

- Группа с состоянием степени тяжести 2. Это самое лёгкое состояние, оно соответствует незначительной травме и не представляет угрозы для жизни пострадавшего.
- Группа с состоянием степени тяжести 3. Такое состояние наступает при серьёзной травме, но не представляет угрозы для жизни пациента.
- Группа с состоянием степени тяжести 4. Это тяжёлое состояние, вызванное серьезными (часто множественными) травматическими повреждениями и угрожающее жизни пациента.
- Группа с состоянием степени тяжести 5. Это самое тяжёлое состояние, называемое критическим или терминальным. Такое состояние наступает при серьёзных травмах, выживание при которых маловероятно.

При разработке методов оценки тяжести состояния пациентов с ТПЖ лучшие результаты получались при объединении групп, соответствующих степеней тяжести 2 и 3, в одну. Таким образом, классификация пациентов с ТПЖ осуществлялась в три группы, характеризующиеся состояниями без угрозы для жизни, тяжёлыми состояниями с угрозой для жизни и критическими состояниями с сомнительным выживанием. Распределение пациентов в группы, в зависимости от степени тяжести состояния, приведено в табл. 4.1. Травмы, соответствующие степени тяжести менее 4, мы в дальнейшем будем условно называть «лёгкими», имея в виду, что они не представляют угрозы для жизни; травмы, соответствующие степеням тяжести 4 и 5 (т.е. с угрозой для жизни пациента), в дальнейшем будем условно называть «тяжёлыми».

Таблица 4.1

Распределение пациентов в группы в зависимости от степени тяжести состояния

Степень тяжести состояния	Группы	Количество пациентов	Всего
Состояния без угрозы для жизни («лёгкие»)	2 и 3	87	87
Состояния с угрозой для жизни («тяжёлые»)	4	74	145
	5	71	

В выборке, использованной для построения математических моделей для прогнозирования исхода ТПЖ (табл. 4.1), было 232 пациента, из которых у 142 исход заболевания был благоприятный, у 90 – летальный.

При срочной оценке тяжести состояния вследствие полученных повреждений использовались следующие показатели: возраст и пол пострадавшего, группа крови и резус-фактор, частота пульса, систолическое и диастолическое артериальное давление, шоковый индекс Альговера (отношение частоты пульса к систолическому давлению), параметры механического воздействия – скорость и сила удара, обстоятельства травмы, а также наличие сочетанных и комбинированных повреждений. Также в качестве потенциальных предикторных признаков рассматривались данные лабораторных анализов. К ним относятся 10 показателей клинического анализа крови: количество эритроцитов, лейкоцитов, эозинофилов, лимфоцитов, моноцитов, палочкоядерных и сегментоядерных нейтрофилов в мазке крови, гемоглобин, СОЭ и цветной показатель

(ЦП), определяемый как отношение утроенного содержания гемоглобина к количеству эритроцитов; 10 показателей биохимического анализа крови: концентрация общего белка, мочевины, креатинина, альфа амилазы, общего, прямого и непрямого билирубина, глюкозы в крови, аспарагиновой трансаминазы (АСТ) и аланиновой трансаминазы (АЛТ); 6 показателей клинического анализа мочи: цвет, концентрация белка, наличие сахара (глюкозы), количество эритроцитов и лейкоцитов в моче, наличие слизи; 4 показателя коагулограммы: время свёртываемости, время рекальцификации цитратной плазмы, потребление протромбина и фибриноген.

Оценка значимости влияния каждого из упомянутых выше показателей на исход и тяжесть состояния пациентов с ТПЖ проводилась с помощью различных статистических критериев в зависимости от шкалы его измерения и закона распределения. Однако, учитывая небольшой размер некоторых групп, для всех показателей в дополнение к параметрическим тестам выявления значимых различий использовались также их непараметрические аналоги. В результате были определены показатели, значимо различающиеся в зависимости от исхода и степени тяжести состояния, а также оценены их пороговые значения, отвечающие каждому из рассматриваемых классов.

4.1.2 Математическая модель прогнозирования исхода ТПЖ

В соответствии с процедурой, описанной в параграфе 3.1, набор категорий значений показателей (табл. 4.2) рассматривался в качестве набора качественных переменных, которые совместно с выходным показателем «исход» были спроектированы в пространство

редуцированной размерности с помощью методов корреспондентского анализа. Размерность пространства выбиралась таким образом, чтобы максимизировать долю объяснённой инерции между 15-тью проецируемыми категориями, в то же время максимально упростив модель. Собственные значения матрицы связей между исследуемыми категориями, процент инерции, объяснённой каждым измерением, кумулятивная инерция, а также график убывания собственных значений в зависимости от увеличения размерности пространства (т.н. график каменистой осыпи) приведены на рис. 4.1.

Таблица 4.2

Области значений показателей–предикторов исхода

Показатель	Исход	
	Благоприятный	Летальный
Диастолическое АД (мм. рт. ст.)	не меньше 65 (≥ 65)	до 65 (40—64, <40)
Шоковый индекс	до 0,8 (≤ 0,8)	более 1,25 (> 1,25, 0,8—1,25)
Возраст (лет)	до 31 (≤ 31)	от 58 (≥ 58, 32—57)
Количество сочетанных травм	нет или 1–2	более 2
Обстоятельства травмы	другое	падение с высоты

Из данных на рис. 4.1. следует, что для объяснения 100% инерции взаимосвязей между 15-тью категориями достаточно девяти измерений. Для получения проекции качества около 90% необходимо

пространство размерности 7. По критерию Кэттеля можно было ограничиться и пространством размерности 5, но кумулятивная инерция, объяснённая пятью измерениями, составляет всего лишь 71,5%. В силу всего вышесказанного было решено использовать проекцию в пространство шести измерений, в котором объясняется 80,8% общей инерции (взаимосвязей между категориями в исходном пространстве).

а)

Eigenvalues and Inertia for all Dimensions (last_Spreadsheet7_(Recovered).sta)
Input Table (Rows x Columns): 15 x 15 (Burt Table)
Total Inertia=1,5000

Number of Dims.	Singular Values	Eigen-Values	Perc. of Inertia	Cumulatv Percent	Chi Squares
1	0,575119	0,330762	22,05080	22,0508	507,0847
2	0,462356	0,213773	14,25156	36,3024	327,7318
3	0,444799	0,197846	13,18974	49,4921	303,3141
4	0,429120	0,184144	12,27624	61,7683	282,3071
5	0,381839	0,145801	9,72009	71,4884	223,5251
6	0,373776	0,139709	9,31391	80,8023	214,1847
7	0,358323	0,128395	8,55968	89,3620	196,8403
8	0,304962	0,093002	6,20011	95,5621	142,5790
9	0,258008	0,066568	4,43787	100,0000	102,0541

б)

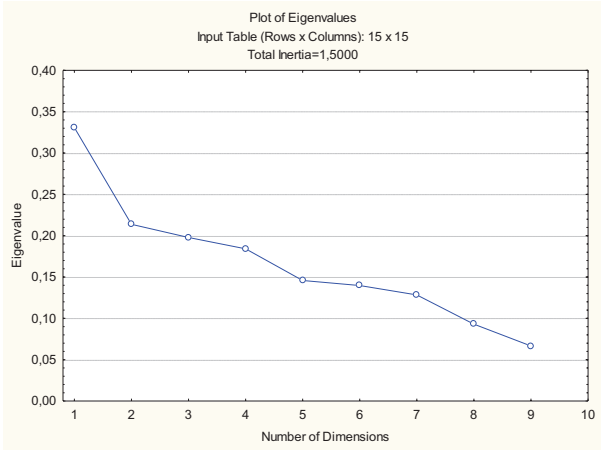


Рис. 4.1. Собственные значения, инерция (а) и график каменистой осыпи (б) для выбора размерности пространства представления предикторов исхода ТПЖ

График проекции признаков-предикторов и класса-исхода в пространство, определяемое первыми двумя собственными значениями, показан на рис. 4.2.

На основании полученного представления в 6-тимерном координатном пространстве были вычислены расстояния между точками-предикторами исхода и точками-исходами. При вычислении расстояний использована обобщённая метрика Минковского: $\rho(x,y)=\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{1/q}$, где показатели степени были выбраны в результате численных экспериментов для обеспечения наилучшей точности: $p=2, q=1,\ \rho(x,y)=\sum_{i=1}^{6}|x_i-y_i|^2$.

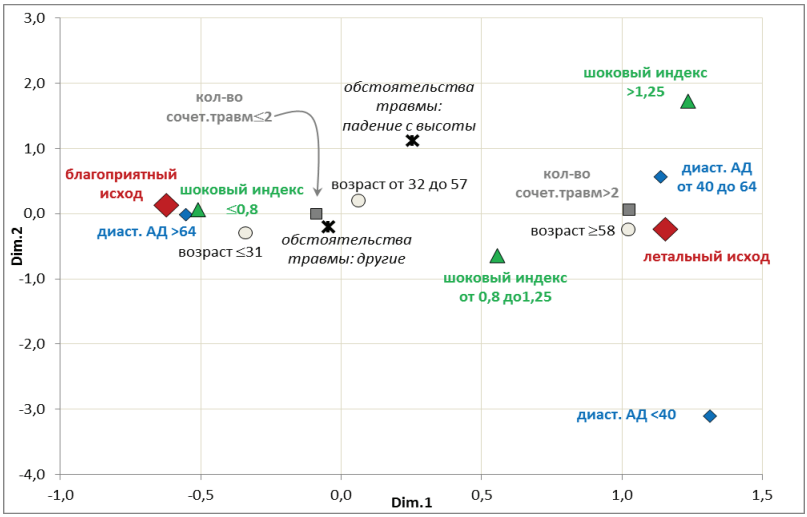


Рис. 4.2. Проекция признаков-предикторов исхода в пространство, определяемое собственными значениями 1 и 2

Весовые коэффициенты предикторов для определения летального и благоприятного исходов вычислены по формуле (3.4), где c_j $(j = 1, 2)$ – прогнозируемый класс, c_1 соответствует благоприятному исходу, c_2 – летальному; υ_l – категории значений пяти используемых предикторов исхода, количество которых в данном случае $\sum_{i=1}^{n} k_i = 13$. Данные весовые коэффициенты приведены в табл. 4.3. Заметим, что т.к. в данном случае решалась задача классификации на два класса, то для облегчения вычислительных процедур по каждому предиктору рассматривалась разность его весовых коэффициентов (приведена в последней колонке табл. 4.3).

Таким образом, получена математическая модель прогнозирования летального исхода при ТПЖ, которую формально можно представить в следующем виде:

$$I_{leth} = \alpha_{\overline{1,3}} \cdot \widetilde{BP} + \alpha_{\overline{4,6}} \cdot \widetilde{IS} + \alpha_{\overline{7,9}} \cdot \widetilde{Age} + \alpha_{\overline{10,11}} \cdot \widetilde{TQ} + \alpha_{\overline{12,13}} \cdot Circ,$$

где BP – диастолическое артериальное давление (мм. рт. ст.), IS – индекс шока Альговера, Age – возраст пациента (лет), TQ – количество сочетанных травм, $Circ$ – обстоятельства травмы, $\widetilde{\odot}$ – категоризированный дихотомический показатель. Значения коэффициентов α_i выбираются в зависимости от значений показателя из табл. 4.3. Т.е. в другой эквивалентной записи I_{leth} представляет собой линейную комбинацию 13-ти дихотомических переменных:

$$I_{leth} = \sum_{i=1}^{13} \alpha_i \cdot x_i ,$$

где: $x_1 = \begin{cases} 1, & BP < 40 \\ 0, & BP \geq 40 \end{cases}$, $x_2 = \begin{cases} 1, & 40 \leq BP \leq 64 \\ 0, & BP > 64 \vee BP < 40 \end{cases}$,

$x_3 = \begin{cases} 1, & BP \geq 65 \\ 0, & BP < 65 \end{cases}$;

$x_4 = \begin{cases} 1, & IS \leq 0{,}8 \\ 0, & IS > 0{,}8 \end{cases}$, $x_5 = \begin{cases} 1, & 0{,}8 < IS \leq 1{,}25 \\ 0, & IS \leq 0{,}8 \vee IS > 1{,}25 \end{cases}$,

$x_6 = \begin{cases} 1, & IS > 1{,}25 \\ 0, & IS \leq 1{,}25 \end{cases}$;

$x_7 = \begin{cases} 1, & Age \leq 31 \\ 0, & Age > 31 \end{cases}$, $x_8 = \begin{cases} 1, & 32 \leq Age \leq 57 \\ 0, & Age \leq 32 \vee Age > 57 \end{cases}$,

$x_9 = \begin{cases} 1, & Age > 57 \\ 0, & Age \leq 57 \end{cases}$;

$x_{10} = \begin{cases} 1, & TQ > 2 \\ 0, & TQ \leq 2 \end{cases}$, $x_{11} = \begin{cases} 1, & TQ \leq 2 \\ 0, & TQ > 2 \end{cases}$;

$x_{12} = \begin{cases} 1, & Circ = \text{падение с высоты} \\ 0, & Circ = \text{другое} \end{cases}$,

$x_{13} = \begin{cases} 0, & Circ = \text{падение с высоты} \\ 1, & Circ = \text{другое} \end{cases}$.

В соответствии с построенной математической моделью при вычисленном значении $I_{leth} > 0$ следует прогнозировать летальный исход, при $I_{leth} < 0$ – бóльшую вероятность благоприятного исхода.

Алгоритм применения математической модели для прогнозирования исхода показан на рис. 4.3.

Таблица 4.3

Весовые коэффициенты предикторов исхода

№ п/п (i)	Значения показателя	Исход		Коэффициент в модели (α_i)
		Благоприятный (ω_{i1})	Летальный (ω_{i2})	
1	Диастолическое АД меньше 40 мм. рт. ст.	0,001	0,016	0,015
2	Диастолическое АД от 40 до 64 мм. рт. ст.	0,006	0,236	0,231
3	Диастолическое АД от 65 мм. рт. ст.	0,739	0,056	–0,683
4	Шоковый индекс не более 0,8	0,099	0,055	–0,044
5	Шоковый индекс от 0,8 до 1,25	0,008	0,146	0,138
6	Шоковый индекс более 1,25	0,003	0,038	0,035
7	Возраст до 31 года	0,016	0,051	0,035
8	Возраст 32–57 лет	0,025	0,098	0,073
9	Возраст больше 57 лет	0,002	0,029	0,026
10	Количество сочетанных травм более 2	0,002	0,024	0,022
11	Количество сочетанных травм нет или 1–2	0,055	0,101	0,046
12	Обстоятельства травмы – падение с высоты	0,006	0,037	0,031
13	Обстоятельства травмы – другое	0,038	0,112	0,074

В результате применения полученной модели к выборке из 216 пациентов с ТПЖ верно было спрогнозировано 122 благоприятных исхода из 142, что соответствует 85,9% специфичности классификатора; и 65 из 74 летальных исходов, что соответствует 87,8% чувствительности. Таким образом, средняя точность прогноза в построенной модели равна 86,9%.

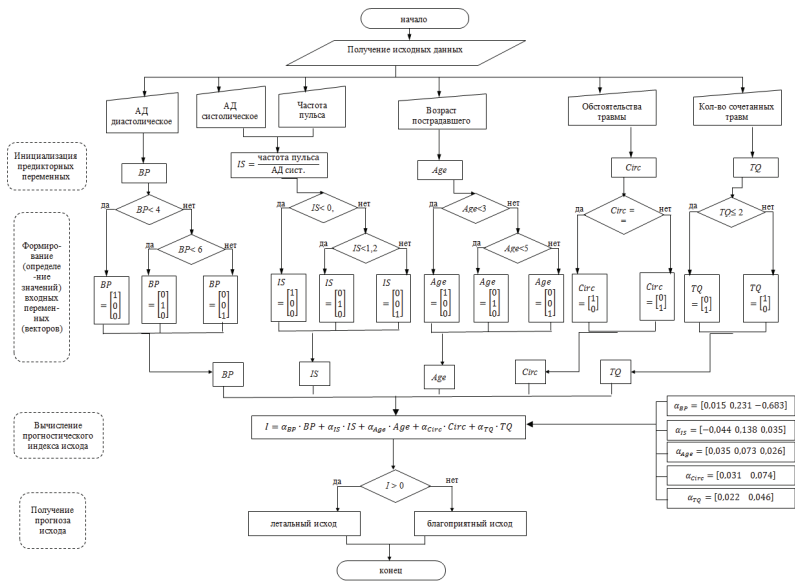


Рис. 4.3. Блок-схема процесса получения прогноза исхода при ТПЖ в соответствии с построенной математической моделью

4.1.3 Математические модели классификации пациентов с ТПЖ по степени тяжести состояния

Для выявления показателей, значимо различающихся в зависимости от степени тяжести состояния пациента при ТПЖ, использовался анализ Краскала—Уоллиса с последующими

попарными сравнениями для значений количественных показателей, и проводился анализ таблиц сопряжённости для качественных переменных. Сам выходной показатель – «степень тяжести» – рассматривался на трёх уровнях градации: степени тяжести менее 3 (“<= 3”), 4 и 5. Проведенные исследования позволили сформировать диапазоны значений предикторов, характерные для различных степеней тяжести (табл. 4.4).

Таблица 4.4

Области значений показателей–предикторов степени тяжести состояния

Показатель	Степень тяжести		
	“<=3”	“4”	“5”
Возраст (лет)	< 35		> 44—45
Количество сочетанных травм	нет (0)	> 1	
Гемоглобин (г/л)	< 104—105, > 104—105		< 104—105
Палочкоядерные нейтрофилы (%)	< 8	> 8	
Сегментоядерные нейтрофилы (%)	> 70—71, < 70—71	> 70—71	< 70—71
Лейкоциты (ед.×10^7/л)	< 12	> 12	
Лимфоциты (%)	Лимфоциты менее 10 – скорее всего «4»-я степень; от 10 до 14 – «4»-я или «5»-я; от 14 до 19 – «5»-я или «<=3»-я; свыше 19 – скорее всего «<=3»-я.		
Моноциты (%)	> 5		> 5
Диастолическое АД (мм. рт. ст.)	>= 70		< 70

Таблица 4.4 (продолжение)

Показатель	Степень тяжести		
	“<=3”	“4”	“5”
Шоковый индекс	< 0,9		>= 0,9
Слизь в моче (увелич. = 1/умерен. = 0)		Если слизь увеличена, то скорее «4», чем «<=3» или «5»	
Обстоятельства травмы (избиение = 3/ другие обстоятельства = 0)	Любые		Скорее всего, НЕ избиение

На следующем этапе было построено три одноклассовых классификатора, каждый из которых позволяет выделять одну из степеней тяжести из всего массива данных.

Для определения состояний каждой отдельной степени тяжести наиболее информативными оказывались различные наборы показателей и категорий их значений (табл. 4.5). Напротив каждой категории показателя в таблице приведен его весовой коэффициент в математической модели, который можно рассматривать как силу влияния этой категории значений показателя на принадлежность объекта к рассматриваемому классу.

Таблица 4.5

Информативные предикторы и их весовые коэффициенты ($\alpha_{k,i}$) для различных степеней тяжести состояния при ТПЖ

№ п/п (i)	Значения показателя	Степень тяжести состояния (*коэффициент модели*)		
		"<= 3" ($\alpha_{3,i}$)	"4" ($\alpha_{4,i}$)	"5" ($\alpha_{5,i}$)
1	Количество сочетанных травм: нет (0)	0,314		
2	Количество сочетанных травм: больше 1	–0,288		
3	Уровень палочкоядерных нейтрофилов: не более 8%	0,058		
4	Уровень палочкоядерных нейтрофилов: больше 8%	–0,076		
5	Уровень лимфоцитов: меньше 10%	–0,009	0,137	
6	Уровень лимфоцитов: от 10% до 14%	–0,020	0,097	
7	Уровень лимфоцитов: от 14% до 19%	0,013	0,039	
8	Уровень лимфоцитов: не менее 19%	0,008	–0,075	
9	Содержание слизи в моче: умеренное		–0,078	
10	Содержание слизи в моче: увеличенное		0,081	
11	Возраст: до 35 лет			–0,041
12	Возраст: 46 лет и старше			0,051
13	Гемоглобин: не больше 105 г/л			0,044
14	Гемоглобин: от 105 г/л			–0,041

Таблица 4.5 (продолжение)

№ п/п (i)	Значения показателя	Степень тяжести состояния (*коэффициент модели*)		
		"<= 3" ($\alpha_{3,i}$)	"4" ($\alpha_{4,i}$)	"5" ($\alpha_{5,i}$)
15	Диастолическое АД: меньше 70 мм. рт. ст.			0,262
16	Диастолическое АД: от 70 мм. рт. ст.			–0,277
17	Шоковый индекс: меньше 0,9			–0,077
18	Шоковый индекс: 0,9 и больше			0,080

Одноклассовый классификатор для определения степени тяжести "<= 3" построен на основании проекции 8-ми категорий предикторов и двух точек, отвечающих за принадлежность и непринадлежность объекта к классу степени тяжести "<= 3", (всего 10 точек) в обобщённое пространство размерности 5. Размерность выбрана на основании анализа собственных значений матрицы сходств таким образом, чтобы обеспечить не менее 90% общей кумулятивной инерции (рис. 4.4).

Eigenvalues and Inertia for all Dimensions (latest_data.sta)
Input Table (Rows x Columns): 10 x 10 (Burt Table)
Total Inertia=1,5000

Number of Dims.	Singular Values	Eigen-Values	Perc. of Inertia	Cumulatv Percent	Chi Squares
1	0,745984	0,556493	37,09951	37,0995	305,4526
2	0,509314	0,259400	17,29335	54,3929	142,3819
3	0,500001	0,250001	16,66674	71,0596	137,2228
4	0,470845	0,221695	14,77969	85,8393	121,6861
5	0,383991	0,147449	9,82995	95,6692	80,9333
6	0,254875	0,064961	4,33075	100,0000	35,6565

Рис. 4.4. Собственные значения и инерция, использованные для выбора размерности пространства представления предикторов степени тяжести "<= 3"

График проекции признаков-предикторов и класса-степени тяжести “<= 3” в пространство, определяемое первыми двумя собственными значениями, показан на рис. 4.5.

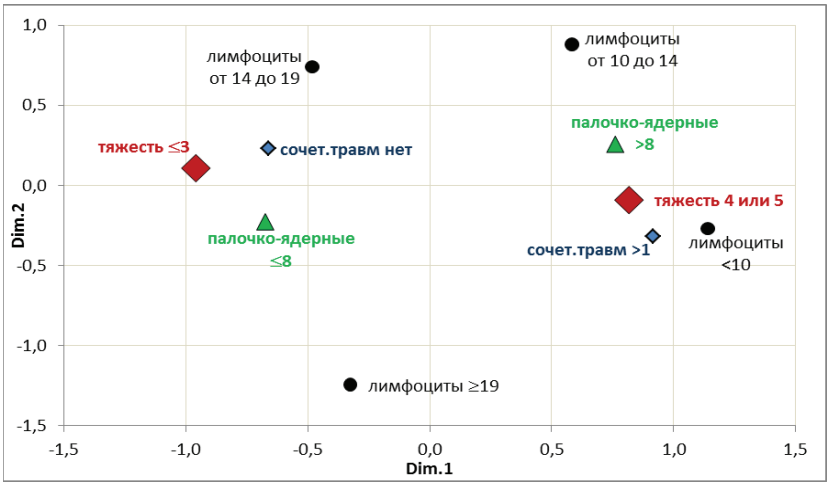


Рис. 4.5. Проекция признаков-предикторов степени тяжести “<= 3” в пространство, определяемое собственными значениями 1 и 2

В построенном 5-тимерном пространстве расстояния точек-предикторов лёгких степеней тяжести до точек, отвечающих прогнозируемому классу, на основании которых получены коэффициенты модели, рассчитаны в метрике Евклида.

Таким образом, получена формула для прогностического индекса лёгких (“<= 3”) степеней тяжести состояния пациентов с ТПЖ:

$$I_{\leq 3} = \alpha_{3,\overline{1,2}} \cdot \widetilde{TQ} + \alpha_{3,\overline{3,4}} \cdot \widetilde{SN} + \alpha_{3,\overline{5,8}} \cdot \widetilde{Lym},$$

где TQ – количество сочетанных травм, SN – уровень палочкоядерных нейтрофилов крови (%), Lym – уровень лимфоцитов крови (%), $\widetilde{\odot}$ – означает категоризированный дихотомический

показатель. Значения коэффициентов $\alpha_{3,i}$ выбираются в зависимости от значений показателя из табл. 4.5. Т.е. формулу индекса для отнесения пациента к лёгкой степени тяжести $I_{\leq 3}$ можно представить в другой эквивалентной записи как линейную комбинацию восьми дихотомических переменных:

$$I_{\leq 3} = \sum_{i=1}^{8} \alpha_{3,i} \cdot x_i \,,$$

где: $x_1 = \begin{cases} 1, & TQ = 0 \\ 0, & TQ > 0 \end{cases}$, $x_2 = \begin{cases} 1, & TQ > 1 \\ 0, & TQ \leq 1 \end{cases}$; $x_3 = \begin{cases} 1, & SN \leq 8 \\ 0, & SN > 8 \end{cases}$,

$x_4 = \begin{cases} 1, & SN > 8 \\ 0, & SN \leq 8 \end{cases}$;

$x_5 = \begin{cases} 1, & Lym < 10 \\ 0, & Lym \geq 10 \end{cases}$, $x_6 = \begin{cases} 1, & Lym \in [10;14) \\ 0, & Lym \notin [10;14) \end{cases}$, $x_7 = \begin{cases} 1, & Lym \in [14;19) \\ 0, & Lym \notin [14;19) \end{cases}$,

$x_8 = \begin{cases} 1, & Lym \geq 19 \\ 0, & Lym < 19 \end{cases}$.

При вычисленном значении $I_{\leq 3} > 0$ степень тяжести состояния пациента оценивается как лёгкая (“<= 3”). Блок-схема алгоритма этого процесса показана на рис. 4.6.

Построенная модель позволяет определять лёгкие степени тяжести состояния с точностью 87,36%.

Для математической модели классификатора, позволяющего определять степень тяжести состояния 4, наиболее информативными (в смысле обеспечения наилучшей точности модели) оказались два показателя – уровень лимфоцитов крови и содержание слизи в моче. Так как содержание лимфоцитов рассматривалось на четырёх уровнях (до 10%, от 10 до 14%, от 14 до 19% и свыше 19%), а содержание

слизи в моче на двух – умеренное и повышенное, размерность общей матрицы сходств (таблицы Бёрта), которая подлежала дальнейшему анализу, равнялась 8. Для получения проекции этой матрицы, 100%-но сохраняющей взаимосвязи между исследуемыми показателями, необходимо пространство из 5-ти измерений, а для достижения качества 90,43% оказалось достаточным четырёхмерного пространства (рис. 4.7).

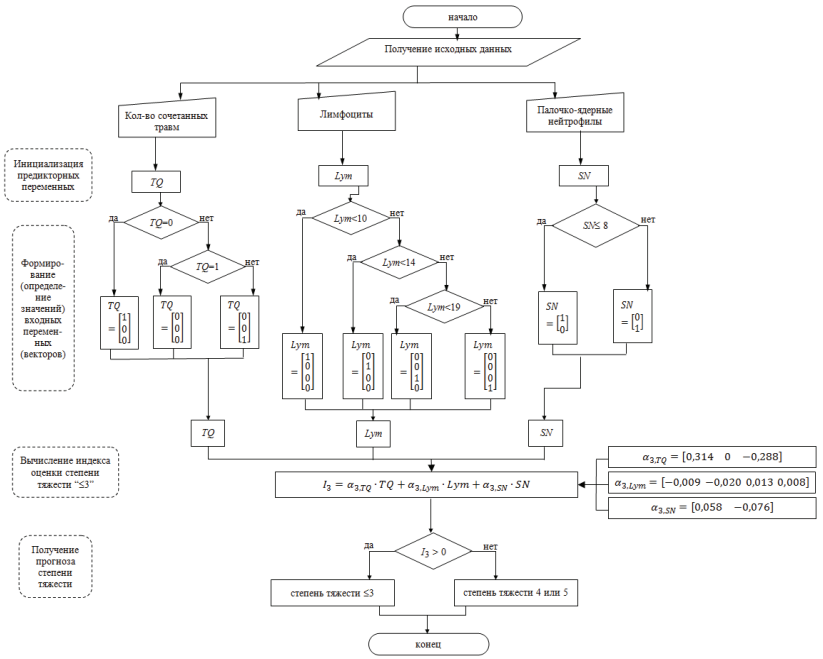


Рис. 4.6. Алгоритм процесса оценки принадлежности пациента к классу степени тяжести состояния "≤ 3"

Графическая интерпретация взаимозависимостей между признаками-предикторами и классом- степенью тяжести "4" в виде

проекции в пространство, определяемое первыми двумя собственными значениями, представлена на рис. 4.8.

	Eigenvalues and Inertia for all Dimensions (latest_data.sta) Input Table (Rows x Columns): 8 x 8 (Burt Table) Total Inertia=1,6667				
Number of Dims.	Singular Values	Eigen-Values	Perc. of Inertia	Cumulatv Percent	Chi Squares
1	0,732606	0,536711	32,20268	32,2027	279,9233
2	0,601346	0,361617	21,69702	53,8997	188,6024
3	0,577350	0,333333	20,00000	73,8997	173,8510
4	0,524832	0,275448	16,52690	90,4266	143,6609
5	0,399445	0,159557	9,57340	100,0000	83,2173

Рис. 4.7. Собственные значения и определяемая ними инерция, используемые для выбора размерности пространства представления предикторов степени тяжести “4”

В полученном 4-хмерном пространстве для вычисления расстояний точек-предикторов степеней тяжести 4 до точек, отвечающих прогнозируемому классу, использовалась Евклидова метрика. На основании значений этих расстояний рассчитаны весовые коэффициенты $\alpha_{4,i}$ модели для определения степени тяжести 4, представленные в табл. 4.5. Сама формула для прогнозирования степени тяжести состояния 4 выглядит следующим образом:

$$I_4 = \alpha_{4,\overline{5,8}} \cdot \widetilde{Lym} + \alpha_{4,\overline{9,10}} \cdot \widetilde{BU},$$

где Lym – уровень лимфоцитов крови (%), BU – уровень содержания слизи в моче, $\widetilde{\odot}$ – означает категоризированный дихотомический показатель. Значения коэффициентов $\alpha_{4,i}$ выбираются в зависимости от значений показателя из табл. 4.5.

Т.е. формулу индекса для отнесения пациента к степени тяжести 4 I_4 можно представить в другой эквивалентной записи как линейную комбинацию шести дихотомических переменных:

$$I_4 = \sum_{i=5}^{10} \alpha_{4,i} \cdot x_i \,,$$

где $$x_9 = \begin{cases} 1, & BU = \text{увеличенное} \\ 0, & BU = \text{умеренное} \end{cases},$$

$$x_{10} = \begin{cases} 0, & BU = \text{увеличенное} \\ 1, & BU = \text{умеренное} \end{cases}.$$

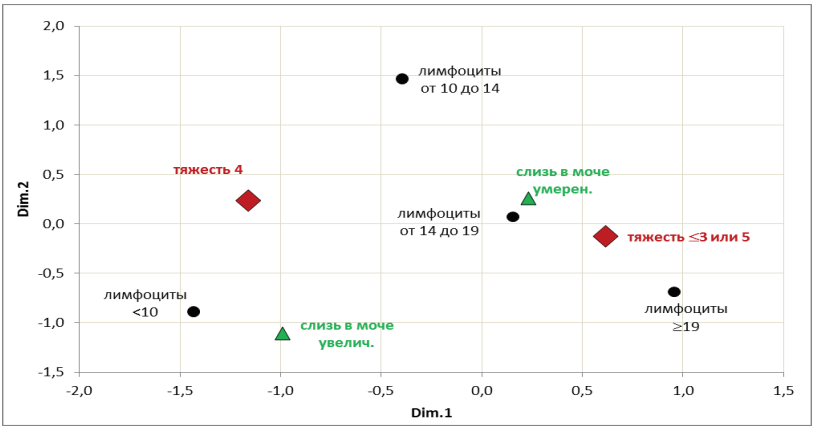


Рис. 4.8. Проекция признаков-предикторов степени тяжести “4” в пространство, определяемое собственными значениями 1 и 2

Схематично алгоритм прогнозирования принадлежности пациента к классу степени тяжести 4 показан на рис. 4.9.

Построенная модель позволяет определять степени тяжести состояния 4 с точностью 83,78%.

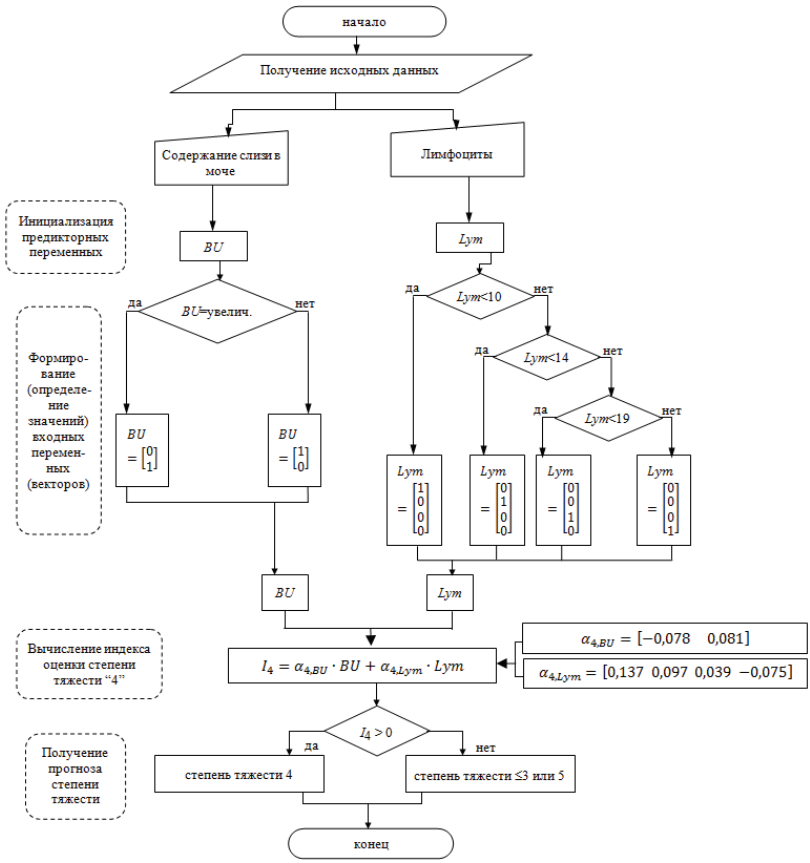


Рис. 4.9. Алгоритм процесса оценки принадлежности пациента к классу степени тяжести состояния "4"

В математической модели классификатора для критических (степени тяжести 5) состояний используются 4 предиктора: возраст пациента, уровень гемоглобина крови, диастолическое АД и индекс шока Альговера. Области значений каждого из предикторов были разделены на 2 категории, таким образом размерность матрицы Бёрта, подаваемой на вход процедуры корреспондентского анализа,

равнялась 10. Данная матрица была спроектирована в пространство, определяемое 4-мя собственными значениями, в котором объясняется 95,38% общей инерции взаимосвязей между предикторами и классом степени тяжести 5 (рис. 4.10).

Number of Dims.	Eigenvalues and Inertia for all Dimensions (latest_data.sta) Input Table (Rows x Columns): 10 x 10 (Burt Table) Total Inertia=1,0000 Singular Values	Eigen-Values	Perc. of Inertia	Cumulatv Percent	Chi Squares
1	0,697706	0,486794	48,67942	48,6794	613,5529
2	0,461787	0,213247	21,32475	70,0042	268,7761
3	0,401073	0,160859	16,08594	86,0901	202,7464
4	0,304862	0,092941	9,29408	95,3842	117,1421
5	0,214844	0,046158	4,61582	100,0000	58,1775

Рис. 4.10. Собственные значения и определяемая ними инерция, используемые для выбора размерности пространства представления предикторов степени тяжести “5”

Графическая интерпретация взаимозависимостей между признаками-предикторами и классом-степенью тяжести “5” в виде проекции в пространство, определяемое первыми двумя собственными значениями, представлена на рис. 4.11. Для вычисления расстояний между точками в четырёхмерном пространстве представления мы пользовались метрикой Евклида.

Математическая модель для прогнозирования критических (степени тяжести 5) состояний формально записывается в виде:

$$I_5 = \alpha_{5,\overline{11,12}} \cdot \widetilde{Age} + \alpha_{5,\overline{13,14}} \cdot \widetilde{Hb} + \alpha_{5,\overline{15,16}} \cdot \widetilde{BP} + \alpha_{5,\overline{17,18}} \cdot \widetilde{IS},$$

где *Age* – возраст пациента (лет), *Hb* –уровень гемоглобина крови (г/л), *BP* –диастолическое артериальное давление (мм. рт. ст.), *IS* – значение индекса шока; $\widetilde{\odot}$ – означает категоризированный дихотомический показатель. Значения коэффициентов $\alpha_{5,i}$ выбираются в зависимости от значений показателя из табл. 4.5. В

эквивалентной записи формула для отнесения пациента к степени тяжести 5 представляется в виде линейной комбинации восьми дихотомических переменных:

$$I_5 = \sum_{i=11}^{18} \alpha_{5,i} \cdot x_i \,,$$

где $x_{11} = \begin{cases} 1, & Age \le 35 \\ 0, & Age > 35 \end{cases}$, $x_{12} = \begin{cases} 1, & Age \ge 46 \\ 0, & Age < 46 \end{cases}$; $x_{13} = \begin{cases} 1, & Hb \le 105 \\ 0, & Hb > 105 \end{cases}$,

$x_{14} = \begin{cases} 0, & Hb \le 105 \\ 1, & Hb > 105 \end{cases}$;

$x_{15} = \begin{cases} 1, & BP < 70 \\ 0, & BP \ge 70 \end{cases}$, $x_{16} = \begin{cases} 0, & BP < 70 \\ 1, & BP \ge 70 \end{cases}$; $x_{17} = \begin{cases} 1, & IS < 0{,}9 \\ 0, & IS \ge 0{,}9 \end{cases}$,

$x_{18} = \begin{cases} 0, & IS < 0{,}9 \\ 1, & IS \ge 0{,}9 \end{cases}$.

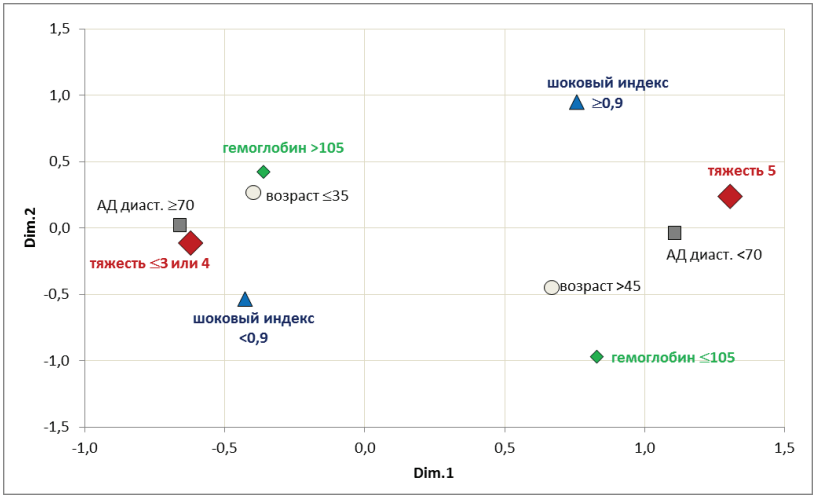


Рис. 4.11. Проекция признаков-предикторов степени тяжести “5” в пространство, определяемое собственными значениями 1 и 2

Данная модель позволяет определять критические степени тяжести состояния с точностью 85,92%. Алгоритм осуществления оценки принадлежности объекта к классу критических состояний приведен на рис. 4.12.

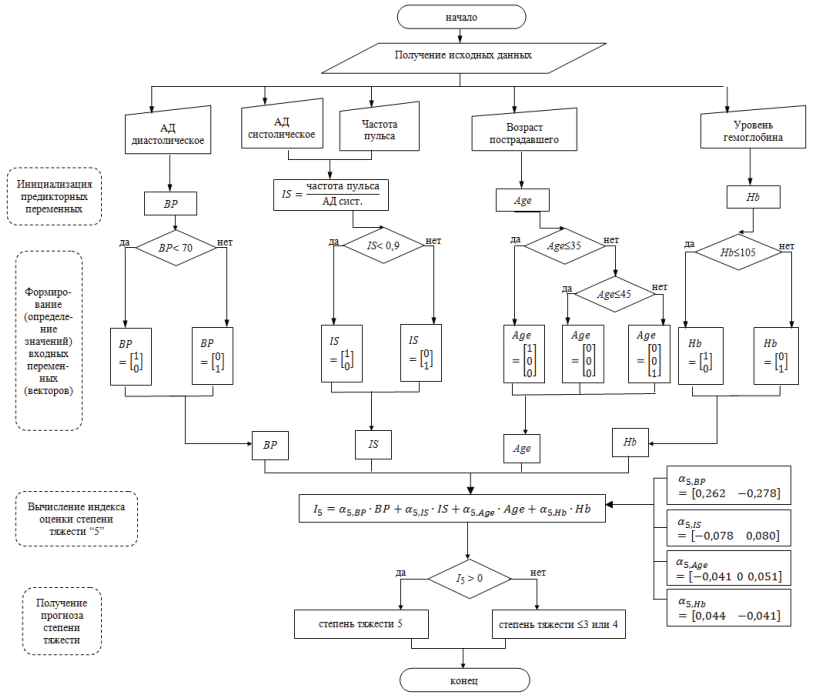


Рис. 4.12. Алгоритм процесса оценки принадлежности пациента к классу степени тяжести состояния “5”

На третьем этапе одноклассовые классификаторы для определения трёх степеней тяжести были объединены в комитет для формирования общей математической модели определения степени тяжести состояния пациентов с ТПЖ. Формирование комитета классификаторов проведено в соответствии с методом рейтингового

голосования по старшинству, процедура которого описана в параграфе 3.2.2 настоящей работы. Исходя из прогностической точности (точности пользователя, user's accuracy) одноклассовых классификаторов I_3, I_4, I_5, оцененной на тестовой выборке и приведенной в табл. 4.6, были определены правила прогнозирования степени тяжести в случае отнесения объекта к более чем одному классу.

Таблица 4.6

Точности пользователя одноклассовых классификаторов для определения трёх классов, соответствующих различным степеням тяжести состояния при ТПЖ

Классификатор (I_k)	Точность пользователя (p_k)	Вероятность ошибки ($1 - p_k$)
I_3	0,62	0,38
I_4	0,40	0,60
I_5	0,71	0,29

Так, при получении трёх положительных ответов одновременно от всех классификаторов рейтинги r_i принадлежности объекта к классам равны:

$$r_{\leq 3} = 0{,}62 + 0{,}60 + 0{,}29 = 1{,}51; \qquad r_4 = 0{,}40 + 0{,}29 + 0{,}38 = 1{,}07;$$

$$r_5 = 0{,}71 + 0{,}38 + 0{,}60 = 1{,}39.$$

Т.е. $r_4 < r_5 < r_{\leq 3}$, следовательно, в этом случае объект следует относить к классу степени тяжести "≤ 3".

При получении положительного ответа одновременно от двух классификаторов имеем:

$$r_{\leq 3} = 0{,}62 + 0{,}60 = 1{,}22 > r_4 = 0{,}40 + 0{,}38 = 0{,}78,$$
$$r_5 = 0{,}71 + 0{,}38 = 1{,}09 > r_{\leq 3} = 0{,}62 + 0{,}29 = 0{,}91,$$
$$r_5 = 0{,}71 + 0{,}60 = 1{,}31 > r_4 = 0{,}40 + 0{,}29 = 0{,}69.$$

Таким образом, если классификаторы $I_{\leq 3}$ и I_4 оба относят объект каждый к своему классу, то следует прогнозировать степень тяжести "≤ 3". Если один из сработавших классификаторов – I_5, то прогнозируемая степень тяжести 5.

Учитывая всё вышесказанное, можем сформулировать общий алгоритм работы комитета классификаторов для оценки степени тяжести состояния пациента при ТПЖ. Сначала необходимо вычислить три индекса оценки для каждой степени тяжести I_3, I_4, I_5. При полученном значении $I_k > 0$ (k = 3,4,5), будем считать, что классификатор I_k выдал положительный ответ, т.е. спрогнозировал «свою» степень тяжести. Затем необходимо подсчитать количество положительных ответов q. При отсутствии положительных ответов (q = 0) комитет отказывается от классификации. Если получен всего лишь один положительный ответ, то объект необходимо отнести к той степени тяжести, классификатор, предназначенный для которой, на нём сработал. Если q = 2 и $I_5 > 0$, то объект относится к классу степени тяжести 5. В случае $I_5 \leq 0$, $I_{\leq 3} > 0$, и $I_4 > 0$, так же, как и при q = 3, необходимо отнести пациента к классу самой лёгкой степени тяжести "≤ 3". Схематично данный алгоритм представлен на рис. 4.13.

Распознавание степени тяжести состояния пациента с ТПЖ с помощью построенного комитета классификаторов позволяет определять лёгкие ("≤ 3") степени тяжести с точностью 88,24%, тяжёлые (степени тяжести 4) состояния – с точностью 83,33%, и критические (степени тяжести 5) состояния – с точностью 89,71%.

Таким образом, в среднем точность классификации составляет 87,09%.

4.1.4 Программная реализация построенных математических моделей

Программное средство для прогноза исхода и оценки тяжести состояния пациентов с ТПЖ и травматическим панкреатитом реализует разработанные в данном разделе работы математические модели и алгоритмы. Программа написана на языке программирования высокого уровня Visual Basic for Applications (VBA) для работы с базами данных форматов *.mdb и *.accdb СУБД Microsoft Office Access 2003—2012.

Функции программы:

- прогнозирование исхода травматического панкреатита и травмы поджелудочной железы на основе математической модели классификации, построенной в параграфе 4.1.2 настоящей работы;
- оценка степени тяжести состояния пострадавших с травмами ПЖ с помощью комитета классификаторов, осуществляющего рейтинговое голосование, и объединяющего три математические модели классификации, описанные в параграфе 4.1.3;
- ввод и сохранение в базе данных сведений о новых пациентах, а также редактирование данных о существующих пациентах;
- дополнительно возможна группировка и отбор данных из БД по различным критериям с просмотром статистики в виде отчетов по имеющейся в базе данных информации с помощью стандартных средств системы управления базами данных Microsoft Office Access 2003—2012.

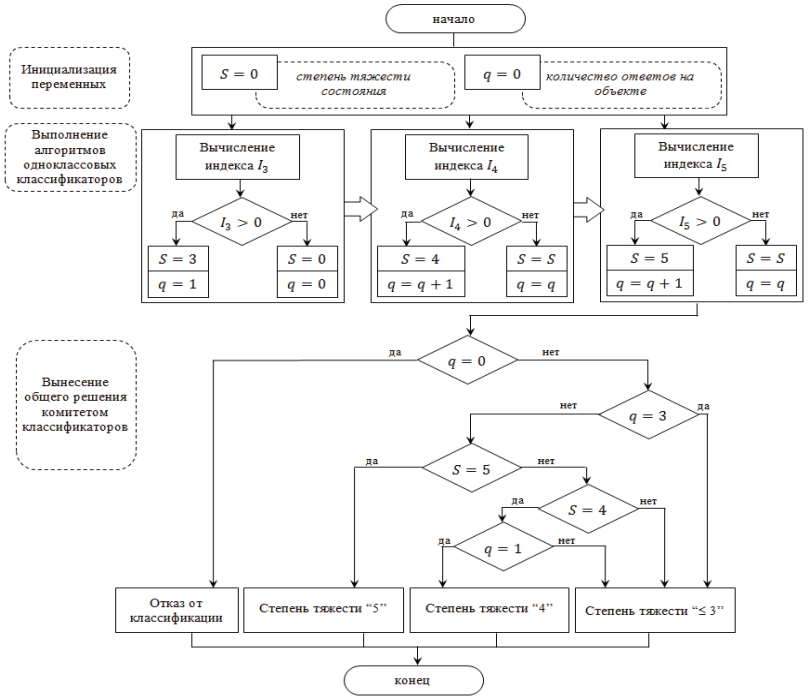


Рис. 4.13. Алгоритм работы комитета классификаторов для оценки степени тяжести состояния при ТПЖ

Основное назначение программы – оценка (прогноз) степени тяжести состояния и исхода при травматических повреждениях поджелудочной железы. При запуске программы открывается главное меню в виде кнопочной формы (рис. 4.14). Работа с программой начинается с выбора действия: для оценки степени тяжести и прогнозирования исхода пользователь может либо выбрать пациента из базы данных, в которой находятся сведения о более чем 200 пациентах с ТПЖ, либо внести данные о новом пациенте.

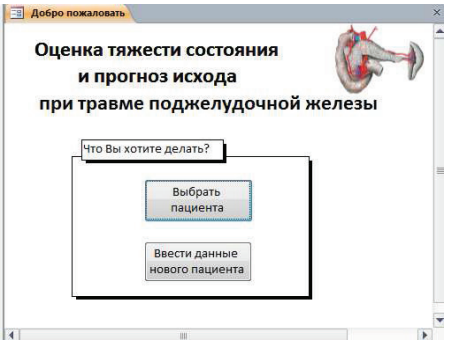


Рис. 4.14. Главное меню (кнопочная форма) программы

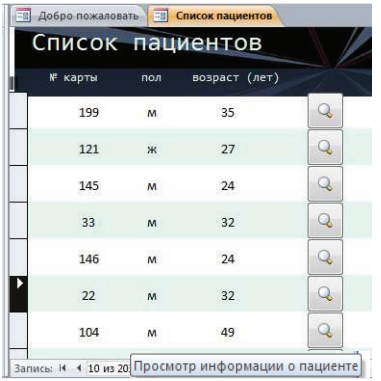


Рис. 4.15. Форма «Выбор пациента»

При нажатии на кнопку «Выбрать пациента» открывается ленточная форма, показанная на рис. 4.15, которая представляет собой алфавитный список пациентов, данные о которых сохранены в базе. По щелчку по кнопке с изображением лупы напротив выбранного пациента возможен просмотр подробной информации о нём, включающей характеристики его физического состояния вследствие травматического повреждения, результаты лабораторных анализов (клинического и биохимического анализа крови, клинического анализа мочи), обстоятельства травмы и др. Эти данные отображаются в форме «Карточка пациента» (рис. 4.16).

По умолчанию форма «Карточка пациента» открывается в режиме просмотра данных. Отредактировать данные можно, нажав на кнопку «Изменить данные пациента» в правом верхнем углу в области заголовка формы. В области примечания внизу формы расположены кнопки для прогнозирования исхода и оценки тяжести состояния пациента. При нажатии на кнопку «Исход» выполняется

подпрограмма для расчёта прогностического индекса исхода ТПЖ, написанная в соответствии с алгоритмом из параграфа 4.1.2, и пользователю выдаётся прогноз исхода для данного пациента в виде текстового сообщения.

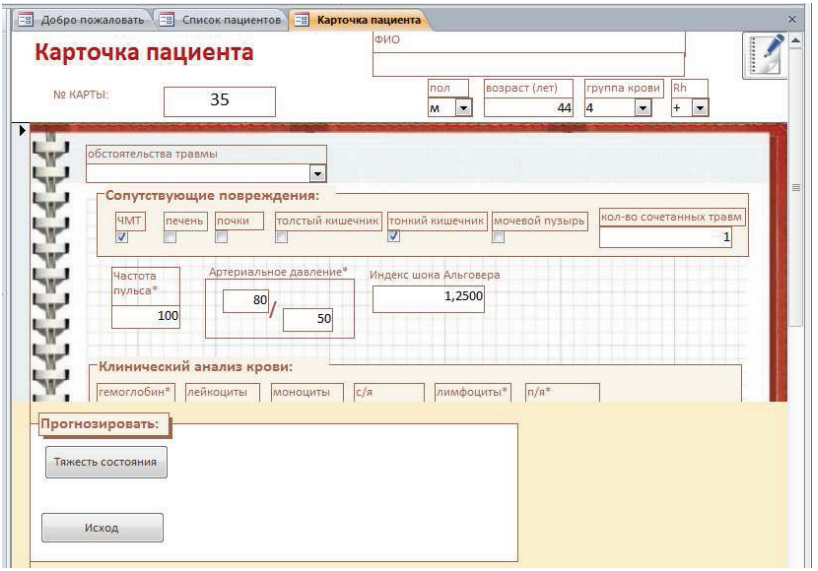


Рис. 4.16. Форма «Карточка пациента» в режиме просмотра данных

По кнопке «Тяжесть состояния» выполняется подпрограмма оценки тяжести состояния, реализованная на основании алгоритмов, описанных в параграфе 4.1.3. Прогнозируемая степень тяжести выдаётся пользователю в виде текстового сообщения: «лёгкое состояние без угрозы для жизни степени тяжести “≤ 3”», «тяжёлое состояние с угрозой для жизни степени тяжести 4» или «критическое состояние с сомнительным выживанием степени тяжести 5».

Для тех пациентов, для которых известен исход их заболевания и (или) априорная оценка степени тяжести состояния, сделанная экспертом (хирургом при поступлении пострадавшего в стационар), пользователь имеет возможность сравнить прогнозные значения степени тяжести состояния с априорными и реальным исходом, нажав кнопку «Сравнить прогноз с реальными данными» внизу примечания формы. Внешний вид формы «Карточка пациента» в этом случае показан на рис. 4.17.

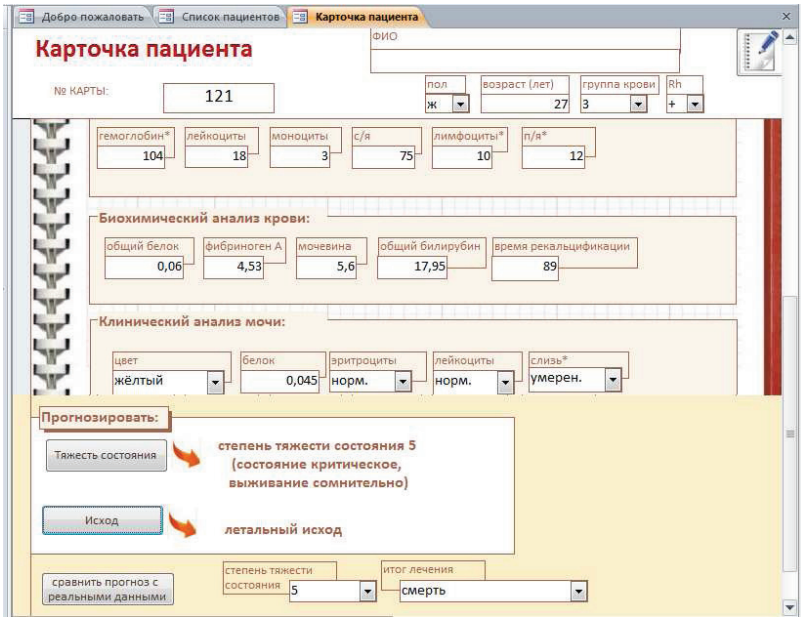


Рис. 4.17. Результат прогнозирования исхода и оценки степени тяжести состояния при ТПЖ
в сравнении с реальными данными для конкретного пациента

4.1.5 Результаты апробации разработанных моделей

При апробации предложенной модели прогнозирования исхода ТПЖ для сравнения была построена модель логит-регрессии для прогнозирования летального исхода:

$$\ln\left(p/(1-p)\right) = 0{,}026 \cdot Age + 0{,}449 \cdot TQ - 0{,}069 \cdot BP + 2{,}102 \cdot IS + 0{,}692,$$

где p – вероятность летального исхода.

Данная модель имела несколько бóльшую специфичность – 90,1%, однако обладала гораздо меньшей чувствительностью – всего 66,7%. Т.е. средняя точность прогноза исхода по модели логит-регрессии составила 78,4%.

Также для сравнения построена модель дерева классификации, использующего тот же набор предикторных показателей для определения клинического исхода ТПЖ. При построении и оптимизации дерева решений использовался метод полного перебора моделей по алгоритму CART с правилом прямой остановки по методу FACT при доле неклассифицированных наблюдений не более 5%. Оптимальное дерево классификации содержало 15 терминальных узлов, его структура показана на рис. 4.18. Дерево решений показало среднюю точность прогнозирования исхода на уровне 82,5%, обладая при этом специфичностью 92,9% и чувствительностью 72,15%.

Сравнительные характеристики точности трёх построенных математических моделей прогнозирования исхода при ТПЖ представлены в табл. 4.7. Как видно, предлагаемая модель на 4,4% превосходит по общей точности модель дерева решений, и на 8,5% модель логит-регрессии. По чувствительности же превосходство разработанной модели перед альтернативными является явно

выраженным: она на 15,6% точнее определяет летальные исходы, чем дерево решений, и на 21,1% точнее, чем модель логит-регрессии.

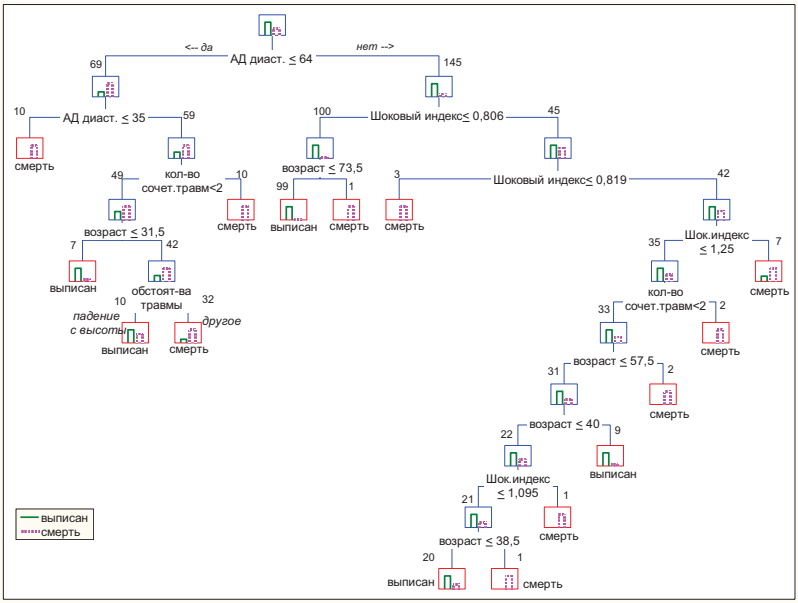


Рис. 4.18. Дерево решений (классификации) для прогнозирования исхода при ТПЖ

При апробации разработанной модели классификации пациентов с ТПЖ на три степени тяжести состояния для сравнения была построена модель с помощью дискриминантного анализа. В качестве входных переменных использовались интервальные показатели из представленных в табл. 4.4. Формирование модели с использованием процедуры пошагового включения переменных позволило получить следующий набор функций классификации для трёх рассматриваемых степеней тяжести состояния:

$$G_{\leq 3} = 0{,}469 \cdot BP + 15{,}708 \cdot IS + 0{,}828 \cdot Lym + 0{,}496 \cdot SN + 0{,}192 \cdot Hb - 45{,}829$$

$$G_4 = 0{,}496 \cdot BP + 14{,}619 \cdot IS + 0{,}608 \cdot Lym + 0{,}664 \cdot SN + 0{,}161 \cdot Hb - 41{,}872$$

$$G_5 = 0{,}281 \cdot BP + 26{,}007 \cdot IS + 0{,}736 \cdot Lym + 0{,}707 \cdot SN + 0{,}152 \cdot Hb - 39{,}794$$

Модель ДА показала общую точность 79,86%, при этом точность распознавания степеней тяжести "≤ 3" составила 82,26%, что на 6% ниже, чем разработанной модели, точность определения состояний степени тяжести 4 – 72,09% (на 11% ниже), точность определения критических состояний (степени тяжести 5) – 84,62% (меньше на 5%). Другие характеристики линейной дискриминантной модели показаны на рис. 4.19.

Таблица 4.7

Сравнительная характеристика точности прогнозирования исхода для трёх математических моделей

Математическая модель	Общая точность (%)	Точность прогнозирования летального исхода (%)
Логит-регрессия	78,4	66,7
Дерево классификации	82,5	72,2
Предлагаемая	86,9	87,8

4.2 Дифференциальная диагностика заболеваний желчевыводящих протоков

4.2.1 Характеристика исходных данных

Для разработки математической модели определения формы заболевания были использованы данные клинических и лабораторных

исследований 90 пациентов, страдающих заболеваниями желчевыводящих протоков различной степени тяжести. В зависимости от формы заболевания все пациенты были разделены на три класса: 1 – пациенты с наиболее лёгкой формой – механической желтухой; 2 – пациенты с диагнозом острый холангит; 3 – пациенты с наиболее тяжёлым диагнозом – билиарный сепсис. Количественный состав классов в наблюдаемой выборке был следующим: 24 пациента относились к классу 1, 46 – к классу 2, и 20 пациентов – к классу 3.

4.2.2 Математическая модель для определения формы заболевания желчевыводящих протоков

Задачи выявления критериев дифференциальной диагностики и построения математической модели классификации пациентов по степени тяжести заболевания решались в несколько этапов. На первом этапе проводилась предобработка данных лабораторных и инструментальных исследований. На втором этапе осуществлялся поиск показателей, обусловливающих различия между введенными нами классами. На третьем этапе проводилось выделение информативных интервалов значений показателей, позволяющих объяснить принадлежность пациента к одному из рассматриваемых классов. На четвёртом этапе определялись весовые коэффициенты каждого из выделенных показателей и формировались решающие правила, позволяющие отнести вновь прибывшего пациента к одной из форм заболевания.

N=144	Discriminant Function Analysis Summary (latest_data.sta) Step 5, N of vars in model: 5; Grouping: Severity (3 grps) Wilks' Lambda: ,22726 approx. F (10,274)=30,076 p<0,0000					
	Wilks' Lambda	Partial Lambda	F-remove (2,137)	p-level	Toler.	1-Toler. (R-Sqr.)
AD-diast	0,350874	0,647696	37,25944	0,000000	0,951037	0,048963
Шоковый индекс	0,314619	0,722332	26,33171	0,000000	0,965526	0,034474
лимфоциты	0,269891	0,842042	12,84988	0,000008	0,945013	0,054987
палочки	0,249989	0,909078	6,85110	0,001459	0,945699	0,054301
гемоглоб	0,246712	0,921155	5,86316	0,003604	0,922438	0,077562

а) Значимость и оценки вкладов переменных в дискриминирующую мощность модели

Group	Classification Matrix (latest_data.sta) Rows: Observed classifications Columns: Predicted classifications			
	Percent Correct	<=3 >=,43056	4 >=,29861	5 >=,27083
<=3	82,25806	51	8	3
4	72,09303	11	31	1
5	84,61539	3	3	33
Total	79,86111	65	42	37

б) Классификационная матрица

Roots Removed	Chi-Square Tests with Successive Roots Removed (latest_					
	Eigen-value	Canonicl R	Wilks' Lambda	Chi-Sqr.	df	p-level
0	1,898584	0,809323	0,227260	205,9510	10	0,000000
1	0,518070	0,584182	0,658731	58,0241	4	0,000000

Variable	Standardized Coefficient for Canonical Variables	
	Root 1	Root 2
AD-diast	0,745689	0,135059
Шоковый индекс	-0,660081	-0,080151
лимфоциты	-0,008785	-0,699743
палочки	-0,199273	0,453328
гемоглоб	0,195711	-0,420649
Eigenval	1,898584	0,518070
Cum.Prop	0,785625	1,000000

в) Значимость, собственные значения и стандартизированные коэффициенты дискриминантных функций

Variable	Factor Structure Matrix (latest_data.sta) Correlations Variables - Canonical Roots (Pooled-within-groups correlations)	
	Root 1	Root 2
AD-diast	0,733584	0,101854
Шоковый индекс	-0,577897	-0,167275
лимфоциты	-0,016133	-0,774763
палочки	-0,131057	0,619806
гемоглоб	0,231247	-0,355942

г) Матрица факторной структуры

Group	Means of Canonical Variables (latest_da	
	Root 1	Root 2
<=3	0,78787	-0,708189
4	0,89205	0,987103
5	-2,23605	0,037495

д) Средние значения канонических переменных

Рис. 4.19. Характеристики модели линейного дискриминантного анализа Фишера для классификации пациентов по степени тяжести состояния при ТПЖ

Для предварительной обработки данных использовались методы описательной статистики и обнаружения выбросов. На этом этапе также исследовались распределения количественных показателей и проверялось их соответствие нормальному закону, для чего использовался критерий Шапиро—Уилка. Поскольку распределения только некоторых из исследуемых показателей в классах подчинялись

нормальному закону, то их дальнейший анализ потребовал применения как параметрических, так и свободных от вида распределения методов статистики.

Для поиска количественных переменных, которыми может быть обусловлено различие между классами, использовался дисперсионный анализ и его непараметрический аналог – анализ Краскала—Уоллиса. Для переменных, значимо различающихся в зависимости от класса, с целью детализации этих различий, проводились попарные сравнения между классами. Для выявления различий между парами классов применялся непараметрический тест независимых выборок Манна—Уитни с поправкой Бонферрони на множественность. Для исследования влияния порядковых признаков на форму заболевания (класс) проводился анализ таблиц сопряжённости. Значимость влияния подтверждалась с помощью критерия $\chi 2$ Пирсона. Все вычисления проводились при доверительной вероятности 95%.

При выделении информативных интервалов значений переменных, объясняющих наличие у пациента той или иной формы заболевания (его принадлежность к классу 1, 2 или 3), определялись пороги показателей, а затем анализировались двухвходовые таблицы сопряжённости формы заболевания (класса) с каждым категоризированным признаком.

На основании проведенных исследований получены статистически обоснованные разбиения количественных показателей на интервалы значений, приведенные ниже.

- Для концентрации непрямого билирубина выделено две категории значений: 1 – до (≤) 40 мкмоль/л (большинство пациентов из 2-го класса характеризуются этими значениями непрямого билирубина) и 2 – свыше 40 мкмоль/л.

- Для показателя АЛТ выделено две категории значений показателя: 1 – до (≤) 1,5 мкмоль/мл·ч и 2 – свыше 1,5 мкмоль/мл·ч (большие значения АЛТ характерны для пациентов из 2-го класса).
- Для протромбинового индекса (ПИ) выделено также две категории значений: 1 – менее 83 % и 2 – от (≥) 83 % (эти значения в бóльшей степени характерны для пациентов из 1-го класса).
- Для концентрации палочкоядерных нейтрофилов выделено два интервала значений: 1 – меньше 5 % (характерно для большинства пациентов из 1-го класса), 2 – от (≥) 5 % (характерно для большинства пациентов из 3-го класса).
- Значения индекса SOFA разделены на две категории: 1 – больше 3 и 2 – до (≤) 3.

Категоризированные признаки и классы (формы заболевания) были спроектированы в единое обобщённое пространство небольшой размерности, где каждый из них представлялся отдельной точкой. Для получения этой проекции, максимально сохраняющей взаимосвязи между переменными, использовался корреспондентский анализ.

Чтобы сохранить не менее 90% взаимосвязей, существующих в исходном пространстве признаков, и достаточно качественно представить 16 имеющихся переменных понадобилось 7 измерений (рис. 4.20).

	Eigenvalues and Inertia for all Dimensions (1-сутки-no-outl Input Table (Rows x Columns): 16 x 16 (Burt Table) Total Inertia=1,2857				
Number of Dims.	Singular Values	Eigen-Values	Perc. of Inertia	Cumulatv Percent	Chi Squares
1	0,57450	0,33005	25,6705	25,6706	253,807
2	0,51653	0,26681	20,7520	46,4226	205,177
3	0,43015	0,18503	14,3913	60,8140	142,288
4	0,36759	0,13512	10,5096	71,3236	103,909
5	0,33837	0,11449	8,9051	80,2288	88,046
6	0,32422	0,10512	8,1760	88,4048	80,837
7	0,27898	0,07783	6,0535	94,4584	59,852
8	0,20860	0,04351	3,3844	97,8428	33,462
9	0,16653	0,02773	2,1571	100,000	21,328

Рис. 4.20. Собственные значения и определяемая ними инерция, используемые для выбора размерности пространства представления признаков, определяющих форму заболевания желчевыводящих протоков

Для примера на рис. 4.21 приведена проекция взаимосвязи между классами-формами заболевания и объясняющими показателями в пространство измерений 1 и 2. Для полного описания взаимосвязей между классами в полученном 7 –ми мерном пространстве можно построить 26 подобных двумерных проекций (карт).

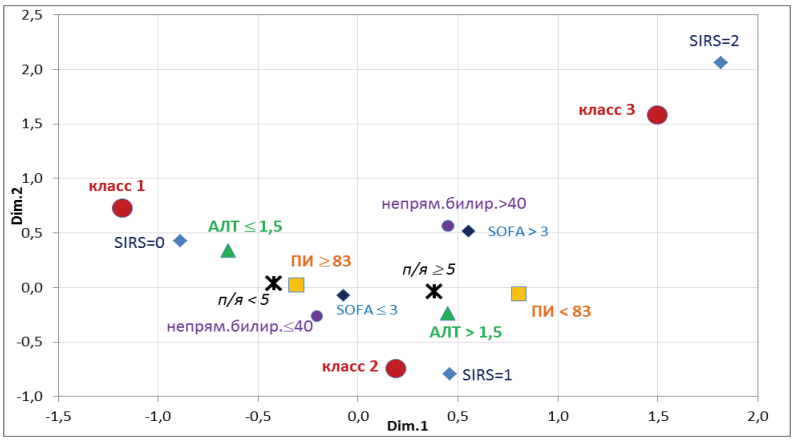


Рис. 4.21. Проекция взаимосвязи между классами и показателями, определяющими форму заболевания желчевыводящих протоков, в пространство измерений 1 и 2

На полученных картах с использованием метрики Евклида были вычислены расстояния от точек-классов (форм заболевания) до точек-признаков (значений показателей). Вес каждого признака в классификации определялся как величина, обратно пропорциональная расстоянию этого признака от соответствующего класса (формы заболевания) в соответствии с формулой (3.4). После вычисления величин весовых коэффициентов выделенных интервалов значений объясняющих показателей в соответствии с формулой (3.5) были построены три оценочные функции (по числу классов).

Для определения класса (формы заболевания), к которому нужно отнести конкретного пациента, необходимо по значениям его показателей для каждого класса вычислить значение его оценочной функции (S_i) по формулам:

Для класса 1 (механическая желтуха):

$$S_1 = 6{,}04 \cdot NCB_1 + 7{,}33 \cdot NCB_2 + 6{,}12 \cdot ALT_1 + 8{,}55 \cdot ALT_2 + 8{,}38 \cdot PI_1 + 4{,}98 \cdot PI_2 + \\ + 8{,}06 \cdot SOFA_1 + 3{,}87 \cdot SOFA_2 + 24{,}34 \cdot SIRS_0 + 5{,}17 \cdot SIRS_1 + 3{,}35 \cdot SIRS_2 + \\ + 6{,}41 \cdot SB_1 + 7{,}39 \cdot SB_2$$

Для класса 2 (острый холангит):

$$S_2 = 4{,}81 \cdot NCB_1 + 10{,}20 \cdot NCB_2 + 10{,}05 \cdot ALT_1 + 5{,}32 \cdot ALT_2 + 8{,}76 \cdot PI_1 + 5{,}02 \cdot PI_2 + \\ + 10{,}32 \cdot SOFA_1 + 3{,}12 \cdot SOFA_2 + 5{,}32 \cdot SIRS_0 + 20{,}50 \cdot SIRS_1 + 2{,}58 \cdot SIRS_2 + \\ + 7{,}38 \cdot SB_1 + 6{,}64 \cdot SB_2$$

Для класса 3 (билиарный сепсис):

$$S_3 = 7{,}82 \cdot NCB_1 + 6{,}86 \cdot NCB_2 + 7{,}34 \cdot ALT_1 + 6{,}84 \cdot ALT_2 + 7{,}04 \cdot PI_1 + 7{,}34 \cdot PI_2 + \\ + 7{,}58 \cdot SOFA_1 + 5{,}20 \cdot SOFA_2 + 6{,}25 \cdot SIRS_0 + 6{,}64 \cdot SIRS_1 + 16{,}80 \cdot SIRS_2 + \\ + 7{,}66 \cdot SB_1 + 6{,}63 \cdot SB_2$$

где $NCB_1 = \begin{cases} 1, & \text{если непрям. билирубин} > 40 \text{ мкмоль/л} \\ 0, & \text{если непрям. билирубин} \le 40 \text{ мкмоль/л} \end{cases}$,

$NCB_2 = 1 - NCB_1$;

$$ALT_1 = \begin{cases} 1, & \text{если АЛТ} > 1{,}5 \text{ мкмоль/мл} \cdot \text{ч} \\ 0, & \text{если АЛТ} \le 1{,}5 \text{ мкмоль/мл} \cdot \text{ч} \end{cases}, \qquad ALT_2 = 1 - ALT_1;$$

$$PI_1 = \begin{cases} 1, & \text{если ПИ} \ge 83\% \\ 0, & \text{если ПИ} < 83\% \end{cases}, \qquad PI_2 = 1 - PI_1;$$

$$SOFA_1 = \begin{cases} 1, & \text{если SOFA} \le 3 \\ 0, & \text{если SOFA} > 3 \end{cases}, \qquad SOFA_2 = 1 - SOFA_1;$$

$$SIRS_0 = \begin{cases} 1, & \text{если SIRS} = 0 \\ 0, & \text{если SIRS} \ne 0 \end{cases}, \qquad SIRS_1 = \begin{cases} 1, & \text{если SIRS} = 1 \\ 0, & \text{если SIRS} \ne 1 \end{cases},$$

$$SIRS_2 = \begin{cases} 1, & \text{если SIRS} = 2 \\ 0, & \text{если SIRS} \ne 2 \end{cases};$$

$$SB_1 = \begin{cases} 1, & \text{если п/я} \ge 5\% \\ 0, & \text{если п/я} < 5\% \end{cases}, \qquad SB_2 = 1 - SB_1.$$

В соответствии с соотношением (3.6) новый пациент должен быть отнесен к классу с наибольшим значением оценочной функции.

На обучающей выборке в первом классе правильно были классифицированы все 24 пациента, во втором – 33 и в третьем – 15. Таким образом, правильно было классифицировано 72 из 90 пациентов, что позволяет говорить от 80%-ной общей точности метода. Более детально точность классификации отражена в таблице 4.8.

4.2.3 Результаты апробации построенной математической модели

Для сравнения была построена модель дискриминантных функций, в которую методом пошагового включения вошли такие показатели: баллы по шкале SIRS, значения непрямого билирубина, протромбиновый индекс (ПИ) и концентрация палочкоядерных нейтрофилов в мазке крови. Для её применения необходимо

вычислить значения трёх функций классификации (F_i) по формулам, приведенным ниже, и отнести пациента к той форме заболевания, для которой получено максимальное значение классификационной функции:

$$F_1 = -18{,}87 \cdot SIRS + 0{,}39 \cdot NCB + 2{,}49 \cdot PI + 2{,}27 \cdot SB - 120{,}69,$$

$$F_2 = -13{,}49 \cdot SIRS + 0{,}32 \cdot NCB + 2{,}42 \cdot PI + 2{,}14 \cdot SB - 113{,}77,$$

$$F_3 = -5{,}03 \cdot SIRS + 0{,}34 \cdot NCB + 2{,}12 \cdot PI + 1{,}58 \cdot SB - 96{,}57,$$

где *SIRS* – баллы пациента по шкале SIRS, *NCB* – значение концентрации непрямого билирубина, *PI* – протромбиновый индекс, и *SB* – концентрация палочкоядерных нейтрофилов в мазке крови.

Таблица 4.8

Точность распознавания формы заболевания желчевыводящих протоков (класса) математической модели, построенной с помощью метода на основании геометрической интерпретации структуры данных

Точность распознавания	Наблюдаемые классы	Спрогнозированные классы		
		1	2	3
Абсолютные значения	1	**24**	0	0
	2	13	**33**	0
	3	2	3	**15**
Точность покрытия классов (%)	1	**100**	0	0
	2	28,3	**71,7**	0
	3	10	15	**75**
Прогностическая точность (%)	1	**61,5**	0	0
	2	33,3	**91,7**	0
	3	5,1	8,3	**100**

Как показывает сравнительный анализ данных таблиц 4.8 и 4.9, модель на основе дискриминантных функций проигрывает по показателям точности определения каждой из форм заболевания модели, построенной на основе авторского метода.

Таблица 4.9

Точность распознавания формы заболевания желчевыводящих протоков (класса) с помощью математической модели, построенной на основе анализа дискриминантных функций

Точность распознавания	Наблюдаемые классы	Спрогнозированные классы		
		1	2	3
Абсолютные значения	1	**23**	1	0
	2	13	**26**	7
	3	2	3	**15**
Точность покрытия классов (%)	1	**95,8**	4,2	0
	2	28,3	**56,5**	15,2
	3	10	15	**75**
Прогностическая точность (%)	1	**60,5**	3,3	0
	2	34,2	**86,7**	31,8
	3	5,3	10	**68,2**

4.3 Определение исхода инсульта

4.3.1 Характеристика исходных данных

Для построения модели использовались данные о 1091 пациенте с ишемическими (ИИ) и геморрагическими инсультами (ГИ), госпитализированном в течение 2008 года в ГКБ № 7 г. Харькова. В исследуемой выборке присутствовали данные о 870 благоприятных

исходах инсульта и о 221 фатальном исходе. Входные показатели для построения классификатора выбирались из совокупности симптомов инсульта, оцениваемых врачом при поступлении пациента в стационар, а также набора факторов риска инсульта, известных из анамнеза. Все эти показатели были измерены в номинальной или, максимум, в порядковой шкале, поэтому основным инструментом их предварительного анализа являлись методы анализа таблиц сопряжённости признаков. В качестве дополнительных параметров использовались балльные оценки самочувствия пациентов, полученные по шкалам GCS (Glasgow Coma Score, шкала комы Глазго) и NIHSS (National Institutes of Health Stroke Scale).

4.3.2 Математическая модель прогнозирования клинического исхода при инсультах

Сначала из общего набора входных переменных с помощью анализа их таблиц сопряжённости с исходом инсульта выделены 46 показателей, статистическая значимость связи которых с исходом инсульта была подтверждена на исследуемой выборке из 1091 человек. Затем отсеяны переменные, отвечающие достаточно редким симптомам (тем, которые наблюдаются в очень малом количестве случаев). По этому принципу, оставлялись показатели, процент наблюдений каждой категории которых был не менее 5% от общего числа случаев. Таким образом были выбраны 27 симптомов и факторов риска. Дополнительно использовались два ранговых показателя – баллы по шкалам GCS и NIHSS. На первоначальным этапе указанные 29 показателей были использованы для построения дерева решений, предназначенного для классификации пациентов в зависимости от клинического исхода инсульта. При исследовании

возможных моделей деревьев классификации наилучшая (рис. 4.22) была получена при использовании для построения дерева метода полного перебора по алгоритму CART с правилом останова – прямая остановка FACT с долей неклассифицированных наблюдений менее 1% от обучающей выборки.

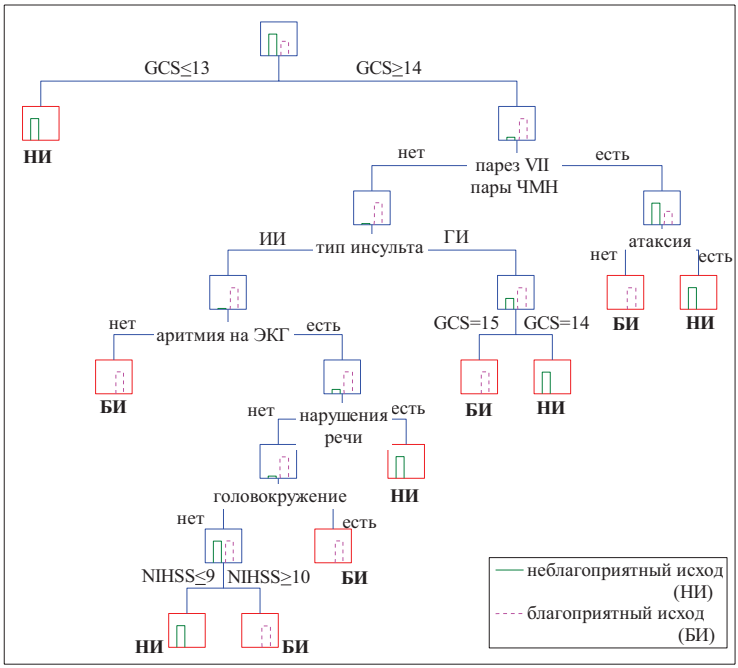


Рис. 4.22. Дерево решений для прогнозирования клинического исхода инсульта

При этом на обучающей выборке (145 пациентов) была получена 100%-но правильная классификация наблюдений. В модель как порядковые предикторы вошли значения баллов по шкалам GCS и NIHSS, как номинальные – тип инсульта, наличие аритмии на ЭКГ, а

также наличие/отсутствие таких симптомов как атаксия, головокружение, ротация языка и парез VII пары черепно-мозгового нерва.

При проверке точности прогнозирования исхода с помощью построенного дерева решений на всей выборке (1091 пациент) были получены следующие характеристики качества классификации:

- общая точность метода составила 85,06%;
- правильно предсказано 717 благоприятных исходов из 870, т.е. точность покрытия (producer's accuracy) класса благоприятных исходов равна 82,41%;
- правильно предсказано 211 летальных исходов из 221, т.е. точность покрытия (producer's accuracy) класса летальных исходов равна 95,48%;
- как благоприятные деревом было классифицировано 727 исходов, из них правильно 98,62%;
- как фатальные деревом было спрогнозировано 364 случая, из них правильно 57,97%.

Таким образом, дерево решений показало довольно приемлемую общую точность и специфичность, и хорошую точность покрытия класса неблагоприятных исходов. При этом, однако, прогностическая точность (user's accuracy) распознавания летальных исходов не достигала и 60%. Это говорит о том, что данная модель имеет существенный недостаток – более 40% случаев, распознаваемых ею как летальные исходы, на самом деле таковыми не являлись.

С целью устранения недостатков дерева решений при прогнозировании исхода инсульта была построена математическая

модель на основе метода, базирующегося на геометрической интерпретации структуры данных, изложенного в параграфе 3.2 настоящей работы. В качестве входных переменных в модели использовались категоризированные значения показателей, участвующие в алгоритме дерева решений. На первом этапе построения модели с помощью корреспондентского анализа было получено графическое представление взаимосвязей входных показателей с исходом инсульта.

Так, на вход процедуры многомерного анализа соответствий подавалась 19×19-матрица Бёрта, кросстабулирующая связи между исходом инсульта и восьмью переменными-предикторами (GCS, NIHSS, тип инсульта, наличие пареза VII пары, ротация языка, атаксия, жалобы на головокружение, наличие аритмии по данным ЭКГ). Следующим шагом была редукция размерности пространства признаков, которая облегчала бы их графическое представление, необходимое для оценки межпризнаковых взаимосвязей, и в то же время максимально сохраняла бы информацию об их сходствах и различиях, присутствующую в исходном пространстве размерности $n = 19$. Выбор оптимальной размерности пространства для представления исследуемых признаков осуществлялся на основании анализа собственных значений матрицы (рис. 4.23). Рассматривались значения размерности от 1 до 10, дающего представление 100%-ного качества. Для достижения же качества проекции не менее 90% оказалось достаточным семи собственных чисел.

Number of Dims.	Singular Values	Eigen-Values	Perc. of Inertia	Cumulatv Percent	Chi Squares
	Eigenvalues and Inertia for all Dimensions (last_Spreadsheet2_(Recovered).sta) Input Table (Rows x Columns): 19 x 19 (Burt Table) Total Inertia=1,1111				
1	0,637085	0,405878	36,52900	36,5290	7676,271
2	0,416619	0,173572	15,62144	52,1504	3282,717
3	0,343424	0,117940	10,61461	62,7651	2230,574
4	0,307174	0,094356	8,49205	71,2571	1784,536
5	0,302191	0,091320	8,21877	79,4759	1727,107
6	0,266324	0,070929	6,38357	85,8594	1341,454
7	**0,233460**	**0,054504**	**4,90532**	**90,7648**	**1030,814**
8	0,211356	0,044671	4,02041	94,7852	844,857
9	0,190143	0,036154	3,25389	98,0391	683,778
10	0,147608	0,021788	1,96093	100,000	412,073

Рис. 4.23. Собственные значения и определяемая ими инерция, используемые для выбора размерности пространства представления признаков, определяющих исход инсульта

Приведенная на рис. 4.24 проекция в пространство измерений 1 и 2, несмотря на то, что её качество представления взаимосвязей всего 52,15%, уже позволяет в первом приближении визуально оценить признаки, характерные для благоприятного и для летального исходов, а также степень влияния каждого симптома на исход.

Полученные координаты точек-симптомов и точек-исходов в 7-ми мерном пространстве были использованы для вычисления весовых коэффициентов предикторов исхода инсульта в формуле прогностического индекса. Весовые коэффициенты вычислялись исходя из принципа: чем ближе точка-симптом находится к точке-исходу, тем больший вес она имеет в прогнозировании данного исхода. Использовалась формула (3.4), в которой расстояние ρ между точками вычислялось в метрике Евклида.

В соответствии с соотношением (3.5) были получены две формулы оценочных функций для благоприятного ($W_{БИ}$) и летального исходов ($W_{НИ}$). Разность $W_{БИ} - W_{НИ} = I$ была названа

нами прогностическим индексом исхода инсульта. Формула для его вычисления получена следующая:

$$I = 0{,}072 \cdot x_1 - 0{,}023 \cdot x_2 - 0{,}332 \cdot x_3 - 0{,}003 \cdot x_4 + 0{,}053 \cdot x_5 + \\ + 0{,}032 \cdot x_6 - 0{,}016 \cdot x_7 + 0{,}011 \cdot x_8 + 0{,}001 \cdot x_9 + 0{,}107 \cdot x_{10} - \\ - 0{,}025 \cdot x_{11} + 0{,}075 \cdot x_{12} - 0{,}017 \cdot x_{13} + 0{,}080 \cdot x_{14} - 0{,}027 \cdot x_{15} + \\ + 0{,}030 \cdot x_{16} - 0{,}009 \cdot x_{17},$$

где x_i $(i = 1,\ldots,17)$ – дихотомические переменные, принимающие значения 0 или 1:

$$x_1 = \begin{cases} 1, \text{при инсульте ишемического типа} \\ 0, \text{при инсульте геморагического типа} \end{cases},$$

$$x_2 = \begin{cases} 1, \text{при инсульте геморагического типа} \\ 0, \text{при инсульте ишемического типа} \end{cases},$$

$$x_3 = \begin{cases} 1, \text{при балле ШКГ} \le 13 \\ 0, \text{при балле ШКГ} > 13 \end{cases}, \qquad x_4 = \begin{cases} 1, \text{при балле ШКГ} = 14 \\ 0, \text{при других баллах ШКГ} \end{cases},$$

$$x_5 = \begin{cases} 1, \text{при балле ШКГ} = 15 \\ 0, \text{при других баллах ШКГ} \end{cases},$$

$$x_6 = \begin{cases} 1, \text{при балле } NIHSS \le 9 \\ 0, \text{при балле } NIHSS \ge 10 \end{cases}, \qquad x_7 = \begin{cases} 1, \text{при балле } NIHSS \ge 10 \\ 0, \text{при балле } NIHSS \le 9 \end{cases},$$

$$x_8 = \begin{cases} 1, \text{при отсутствии жалоб на голокружение} \\ 0, \text{при наличии жалоб на голокружение} \end{cases},$$

$$x_9 = \begin{cases} 1, \text{при наличии жалоб на голокружение} \\ 0, \text{при отсутствии жалоб на голокружение} \end{cases},$$

$$x_{10} = \begin{cases} 1, \text{при отсутствии пареза 7 пары} \\ 0, \text{при наличии пареза 7 пары} \end{cases},$$

$$x_{11} = \begin{cases} 1, \text{при наличии пареза 7 пары} \\ 0, \text{при отсутствии пареза 7 пары} \end{cases},$$

$$x_{12} = \begin{cases} 1, \text{при нормальной ротации языка} \\ 0, \text{в противном случае} \end{cases},$$

$$x_{13} = \begin{cases} 1, \text{при нарушенной ротации языка} \\ 0, \text{в противном случае} \end{cases},$$

$$x_{14} = \begin{cases} 1, \text{при отсутствии атаксии} \\ 0, \text{при наличии атаксии} \end{cases},$$

$$x_{15} = \begin{cases} 1, \text{при наличии атаксии} \\ 0, \text{при отсутствии атаксии} \end{cases},$$

$$x_{16} = \begin{cases} 1, \text{при отсутствии аритмии по данным ЭКГ} \\ 0, \text{при наличии аритмии по данным ЭКГ} \end{cases},$$

$$x_{17} = \begin{cases} 1, \text{при наличии аритмии по данным ЭКГ} \\ 0, \text{при отсутствии аритмии по данным ЭКГ} \end{cases}.$$

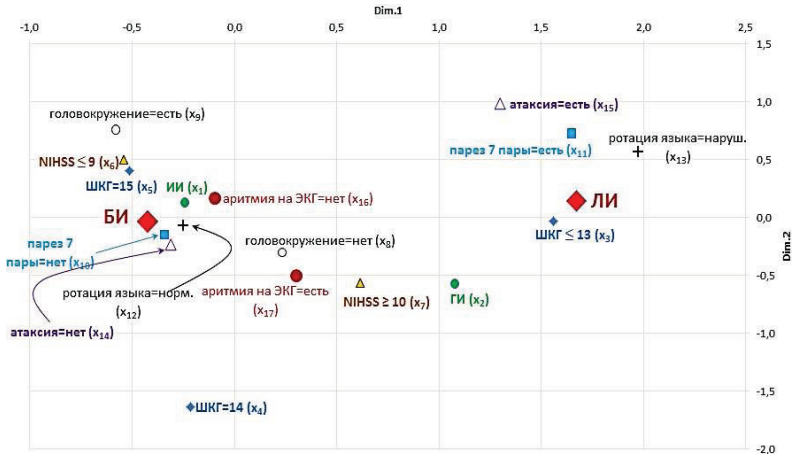


Рис. 4.24. Карта-проекция взаимосвязей признаков, объясняющих исход инсульта, в пространство, определяемое первыми двумя собственными числами

Вычисленное таким образом значение индекса I теоретически может изменяться в пределах от –1 до +1 (практически от –0,46 до +0,46), причём, положительные значения соответствуют благоприятному прогнозу, отрицательные – прогнозу летального (неблагоприятного) исхода инсульта. Значения прогностического индекса, близкие к +0,46, свидетельствуют о высокой вероятности благоприятного исхода инсульта, близкие к –0,46 о большой

вероятности летального исхода. При $I = 0$ следует говорить о равной вероятности обоих исходов, т.е. значения I около 0 получаются в случаях, когда однозначный прогноз исхода сделать затруднительно.

4.3.3 Результаты апробации построенной математической модели

При проверке качества прогнозирования исхода с помощью предлагаемой процедуры вычисления прогностического индекса на всей выборке (1091 пациент) были получены следующие характеристики точности:

- общая точность метода составила 91,93%;
- правильно предсказано 835 благоприятных исходов из 870, т.е. точность покрытия (producer's accuracy) класса благоприятных исходов равна 95,98%;
- правильно предсказано 168 летальных исходов из 221, т.е. точность покрытия (producer's accuracy) класса летальных исходов равна 76,02%;
- как благоприятные с помощью прогностического индекса определены 888 исходов, из них правильно 94,03%;
- как фатальные с помощью прогностического индекса классифицированы 203 наблюдения, из них правильно 82,76%.

Для сравнения показатели точности дерева решений и модели, построенной на базе метода, основанного на геометрической интерпретации структуры данных, при прогнозировании клинического исхода инсульта приведены в таблице 4.10.

Как видим, общая точность модели классификатора на основе геометрического подхода почти на 7% выше, чем у дерева решений. В

то же время она несколько проигрывает дереву решений по точности покрытия (producer's accuracy) класса неблагоприятных исходов. Однако, прогностическая точность (user's accuracy) распознавания летальных исходов у этой модели существенно выше. Более того, соотношение между показателями точности производителя и точности пользователя (producer's / user's accuracy) у этой модели более сбалансированное, что позволяет ожидать меньшего количества ошибок прогнозов.

Таблица 4.10

Точность прогнозирования исхода инсульта с помощью математических моделей на основе деревьев решений и метода, базирующегося на геометрической интерпретации структуры данных

Точность распознавания	Наблюдаемые классы	Спрогнозированные классы			
		Дерево решений		Авторский метод	
		БИ	НИ	БИ	НИ
Абсолютные значения	БИ	**717**	153	**835**	35
	НИ	10	**211**	53	**168**
Точность покрытия классов (%)	БИ	**82,41**	17,59	**95,98**	4,02
	НИ	4,52	**95,48**	23,98	**76,02**
Прогностическая точность (%)	БИ	**98,62**	42,03	**94,03**	17,21
	НИ	1,38	**57,97**	5,97	**82,76**
Общая точность (%)		85,06		91,93	

На основании полученных в данном разделе результатов можно сделать следующие выводы:

1. Математическая модель прогнозирования клинического исхода при травме ПЖ и травматическом панкреатите, построенная с помощью разработанного метода классификации на основе геометрической интерпретации структуры данных, позволяет предсказывать исход с точностью 86,9%, что на 8,5% выше точности модели логит-регрессии и на 4,4% выше точности модели дерева классификации, построенных для решения той же задачи. Построенная на основе авторского метода математическая модель обладает высокой чувствительностью, поскольку точность прогнозирования летального исхода составляет 87,8%, что на 21,1% выше точности модели логит-регрессии и на 15,6% выше точности модели дерева классификации, построенных для решения той же задачи.

2. Посредством разработанной информационной технологии, объединяющей метод построения классификаторов, базирующийся на анализе пространственного представления признаковых переменных, и метод построения ансамбля классификаторов на основе совмещения принципов машины покрывающих множеств и рейтингового голосования, синтезированы математические модели и алгоритмы для распознавания (оценки) степени тяжести состояния пострадавших с ТПЖ, показавшие общую точность 87,3%, что на 7,2% выше точности модели дискриминантных функций, построенной для решения этой же задачи. При этом точность распознавания степени тяжести “≤3” составила 88,2%, что на 6% выше, чем у модели дискриминантных функций; точность определения степени тяжести “4” составила 83,3%, что на 11,2% выше, чем у модели дискриминантных функций; точность распознавания степени тяжести “5” составила 89,7%, что на 5,1% выше, чем у модели дискриминантных функций.

3. Разработанные алгоритмы и построенные математические модели явились методической основой для создания программного обеспечения, ориентированного на медицинский персонал, которое

может быть рекомендовано в качестве практического инструмента для прогнозирования исхода и оценки тяжести состояния пациентов с травмами поджелудочной железы и травматическим панкреатитом.

4. Применение разработанного метода построения классификаторов на основании геометрической интерпретации структуры многомерных данных позволило построить математическую модель дифференциальной диагностики заболеваний желчевыводящих протоков, обладающую общей точностью 80%, что на 9% выше, чем точность модели, построенной для решения этой же задачи с помощью дискриминантного анализа. При этом точность определения наиболее лёгкой формы заболевания (механическая желтуха), оцененная на тестовой выборке, составила 100%, что на 4% выше точности модели дискриминантных функций; точность распознавания острого холангита составила 72%, что на 15% выше точности модели дискриминантных функций; точность определения наиболее тяжёлой формы заболевания (билиарный сепсис) составила 75% и не отличалась от точности модели дискриминантных функций.

5. Математическая модель прогнозирования клинического исхода инсульта, построенная на базе разработанного метода классификации на основании геометрической интерпретации структуры многомерных данных, позволяет определять исход заболевания с общей точностью 91,9%, что на 6,9% выше, чем точность модели дерева решений, построенной для тех же целей. При этом предлагаемая модель имеет высокую чувствительность – 82,8%, которая на 24,8% превышает чувствительность модели дерева решений.

ЗАКЛЮЧЕНИЕ

В монографии предложено новое решение актуальной научно-прикладной задачи – разработки моделей и методов классификации для медицинских приложений и их реализации в информационной технологии поддержки принятия, что позволяет повысить точность дифференциальной диагностики, определения степени тяжести состояния пациента и прогнозирования клинического исхода за счёт использования моделей, построенных на основе разработанных методов классификации с обучением на основании геометрической интерпретации структуры многомерных данных и формирования композиций классификаторов.

1. Проведенный анализ существующих методов построения моделей классификации с обучением позволил выделить как наиболее перспективные в медицинских приложениях методы KDD, не налагающие жёстких требований ни на вид модели, ни на свойства входных данных, среди которых обнаружен класс методов графической интерпретации и визуализации многомерных данных, перспективы и результаты практического применения которых ещё недостаточно изучены. В свою очередь, анализ методов геометрической интерпретации структуры данных позволил выбрать технику множественного корреспондентского анализа как наиболее перспективную для использования при разработке метода построения классификаторов по обучающей информации, т.к. она позволяет работать с признаками, измеренными в наипростейшей шкале – шкале наименований, не налагая дополнительных требований на их типы данных и законы распределения, и получать при этом достаточно

легко интерпретируемые представления взаимосвязей признаков с классами объектов в виде карт в пространстве небольшой размерности.

2. Разработанный метод построения классификаторов по обучающей информации, основанный на геометрической интерпретации структуры многомерных данных и представлении о классе не как о подмножестве пространства объектов, а как о ещё одной составляющей его признакового описания, обеспечивает совместное использование качественных и количественных независимых переменных и даёт возможность создавать модели алгоритмов типа вычисления оценок для классификации пациентов по диагнозу, а также получать решения других задач классификации с обучением, имеющие повышенную точность.

3. Предложенный метод количественной оценки силы влияния описывающих признаков на принадлежность объектов к классам, базирующийся на метрическом подходе к анализу пространственной структуры взаимосвязей между предикторными признаками и классами, позволяет учитывать нелинейность и немонотонность поведения предикторных переменных при переходе от класса к классу, что способствует повышению качества классификации.

4. Предложенный метод составления композиций классификаторов «рейтинговое голосование», являющийся развитием процедур взвешенного голосования и смесей экспертов за счёт совместного учёта не только точности базовых алгоритмов при прогнозировании отдельных классов, а также и их ошибок на других классах, позволяет повысить качество классификации в случаях совместного использования нескольких классификаторов.

Формализованный алгоритм построения композиций

классификаторов «рейтинговое голосование по старшинству» за счёт использования синтеза эвристик, лежащих в основе взвешенного голосования и голосования по старшинству, применение которого целесообразно в случаях, когда базовые классификаторы возможно ранжировать в порядке их точности на различных классах, усовершенствует стандартную процедуру голосования по старшинству, что даёт возможность уменьшить количество отказов от классификации по сравнению со стандартными композициями с логикой старшинства или большинства.

5. Применение построенных моделей и алгоритмов, базирующихся на использовании предложенных метода построения классификаторов на основе анализа пространственного представления признаковых переменных и методов формирования композиций классификаторов на основе совмещения принципов специализации и взвешенного голосования, а также информационной технологии, реализующей их в автоматизированной системе, позволило повысить точность оценки степени тяжести состояния пациентов и предсказания клинического исхода при травме поджелудочной железы и травматическом панкреатите, дифференциальной диагностики заболеваний желчевыводящих протоков и прогнозирования летальности при инсультах.

Разработанные алгоритмы и построенные математические модели явились методической основой для создания программного обеспечения, ориентированного на медицинский персонал, которое может быть рекомендовано в качестве практического инструмента для прогнозирования клинического исхода и оценки тяжести состояния пациентов с травмами поджелудочной железы и травматическим панкреатитом.

6. Проведенный сравнительный анализ выявил, что разработанные на основе предложенного в работе метода построения классификаторов модели обладают бóльшей точностью по сравнению с моделями, построенными на основе классических методов статистического моделирования (анализ дискриминантных функций, логистическая регрессия, метод деревьев решений). По показателям общей точности классификации превышение составляет 4,4÷8,9%, по специфичности – 4,2÷6%, по чувствительности – 7,6÷24,8%.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Ефименко И.В. Интеллектуальные системы поддержки принятия решений в медицине: ретроспективный обзор состояния исследований и разработок и перспективы / И.В. Ефименко, В.Ф. Хорошевский // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2017) : материалы международной научно-технической конференции (Минск, 16 - 18 февраля 2017 года) / редкол. : В. В. Голенков (отв. ред.) [и др.]. – Минск : БГУИР, 2017. – С. 251–260.

2. Гусев А.В. Поддержка принятия врачебных решений в медицинских информационных системах медицинской организации / А.В. Гусев, Т.В. Зарубина // Врач и информационные технологии, 2017, № 2. – С. 60–72.

3. Жаркова О.С. Построение систем поддержки принятия решений в медицине на основе деревьев решений / О.С. Жаркова, К.А. Шаропин, А.С. Сеидова, Е.В. Берестнева, И.А. Осадчая // Современные наукоемкие технологии. – 2016. – № 6-1. – С. 33–37

4. Ахутин В.М. Оценка качества формализованных медицинских документов / В.М. Ахутин, В.В. Шаповалов, М.О. Иоффе // Медицинская техника. – №2. – 2002. – С. 27–31.

5. Поворознюк А.И. Система поддержки принятия решения в медицине на основе синтеза структурированных моделей объектов диагностики / А.И. Поворознюк // Научные ведомости БелГУ. Серия: История. Политология. Экономика. Информатика. 2009. №12–1. — [Электронный ресурс] — Режим доступа: http://cyberleninka.ru/article/ n/sistema-podderzhki-prinyatiya-resheniya-v-meditsine-na-osnove-sinteza-strukturirovannyh -modeley-obektov-diagnostiki (дата обращения: 26.12.2014).

6. Малых В.Л. Управляемый стохастический прецедентный процесс с памятью как математическая модель лечебно-диагностического процесса / В.Л. Малых, Я.И. Гулиев // Информационные технологии и вычислительные системы. — 2014. — № 2. – С. 60–72.

7. Дувалкина А.В. Система поддержки принятия решений в медицине // Современные научные исследования и инновации. 2017. № 2 [Электронный ресурс]. URL: http://web.snauka.ru/issues/2017/02/78010 (дата обращения: 11.01.2018).

8. Халафян А.А. Анализ и синтез медицинских систем поддержки принятия решений на основе технологий статистического моделирования : автореф. дисс. на соискание уч. степени д. техн. наук: 05.13.01 “Системный анализ, управление и обработка информации (по отраслям)”/ Халафян Алексан Альбертович ; Кубанский гос. технологический университет – Краснодар, 2010. – 47 с.

9. Доан Д.Х. Обзор подходов к проблеме принятия решений в медицинских информационных системах в условиях неопределенности / Д.Х. Доан, А.В. Крошилин, С.В. Крошилина // Фундаментальные исследования. – 2015. – № 12-1. – С. 26–30.

10. Симанков В.С. Системный анализ и современные информационные технологии в медицинских системах поддержки принятия решений / В.С. Симанков, А.А. Халафян. – М.: БИНОМ, 2009. – 362 с.

11. Champion H.R., Sacco W.J., Hannan D.S., Lepper R.L., Atzinger E.S., Copes W.S., Prall R.H. Assessment of Injury Severity: the Triage Index. *Critical Care Medicine*, 1980, vol. 8, pp. 201–208.

12. Kaukinen L, Pasanen M, Kaukinen S, Ojanen R. Predicting the prognosis with trauma indices in surgical patients treated in the intensive care unit. *Ann ChirGynaecol*, 1984, no. 73(5), pp. 253–260.

13. Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин) / К.В. Воронцов. – [Электронный ресурс] – Режим доступа: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (дата обращения: 20.12.2014).

14. Журавлёв Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации. // Проблемы кибернетики. – Вып. 33. – 1978. – С. 5–68.

15. Прикладная статистика: Классификация и снижение размерности: Справ.изд. / [Айвазян С.А., Буштабер В.М., Енюков И.С., Мешалкин Л.Д.]. – М.: Финансы и статистика, 1989 – 607 с.

16. Герасимов А.Н. Медицинская статистика : Учебн. Пособие / А.Н. Герасимов – М.: ООО «Медицинское информационное агентство», 2007. – 480 с.

17. Кармазановский Г.Г. Оценка диагностической значимости метода («чувствительность», «специфичность», «общая точность). // Анналы хирургической гепатологии. – Т. 2. – 1997. – С. 139–142.

18. Kuncheva L.I. *Combining pattern classifiers: methods and algorithms*. Hoboken, New Jersey: "A Wiley-Interscience publication." A John Wiley & Sons, Inc., 2004. 360 p.

19. Karimollah Hajian-Tilaki Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation /Caspian J Intern Med. 2013 Spring; 4(2): 627–635.

20. Zweig M.H., Campbell G. ROC Plots: A Fundamental Evaluation Tool in Clinical Medicine // Clinical Chemistry, Vol. 39, No. 4, 1993.

21. Леонов В.П. Основные понятия ROC-анализа. / В.П. Леонов. – [Электронный ресурс] – Режим доступа: http://www.biometrica.tomsk.ru/ROC-analysis.pdf (дата обращения: 20.03.2018).

22. Thomas G. Tape. Interpreting Diagnostic Tests / Tape Th. G. – [Электронный ресурс] – Режим доступа: http://gim.unmc.edu/dxtests (дата обращения: 20.03.2018).

23. Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers // 2004 Kluwer Academic Publishers.

24. Encyclopedia of Machine Learning / Claude Sammut, Geoffrey I, Webb (eds.) / Springer, Boston, MA, 2011.

25. Powers David M.W. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Technical Report SIE-07-001 / School of Informatics and Engineering, Finders University, Adelaide, Australia*, 2007, 24 p.

26. Faecett T. An introduction to ROC analysis. *Pattern Recognition Letters*, 2006, Vol. 27, pp. 861–874.

27. Davis J., Goadrich M. The Relationship Between Precision-Recall and ROC Curves // Proc. Of 23 International Conference on Machine Learning, Pittsburgh, PA, 2006

28. Congalton R.G. A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. *Remote Sens. Environ.*, 1991, No. 37, pp. 35–46.

29. Fisher R.A. The use of multiple measurements in taxonomic problem. *Ann. Eugen*, 1936, No. 7, pp. 179–188.

30. Орлов А.И. Прикладная статистика /Орлов А.И. – М.: “Экзамен”, 2004. – 576 с.

31. Капшитарь А.А. Математическая модель прогноза исхода закрытой травмы печени / А.А. Капшитарь, А.В. Капшитарь, И.Ф. Сырбу // Украинский журнал хирургии. – 2012. – № 1 (16). – С. 61–66.

32. Факторный, дискриминантный и кластерный анализ / Ким Дж.-О., Мьюллер Ч.У., Клекка У.Р., Олдендерфер М.С., Блэшфилд Р.К.;

пер. с англ.; под ред. Енюкова И.С. – М.: Финансы и статистика. – 1989 – 215 с.

33. Использование дискриминантного анализа для диагностики хронической сердечной недостаточности у подростков / Бых А.И., Высоцкая Е.В., Порван А.П. [и др.] // Вестник НТУ «ХПИ». – 2010. – № 3. – С. 16–22.

34. Беркасова И.В. Прогнозирование риска периоперационных осложнений в реконструктивной хирургии пищевода / И.В. Беркасова, Е.И. Верещагин, И.М. Митрофанов // «Медицина и образование в Сибири», № 2, 2013. // Сетевое научное издание Новосибирского Государственного Медицинского Университета. – [Электронный ресурс] – Режим доступа: http://www.ngmu.ru/cozo/mos/article/text_full.php?id=980 (дата обращения: 02.01.2014).

35. Бондарь И.А. Прогнозирование риска развития микрососудистых осложнений при сахарном диабете 1-го типа / И.А. Бондарь, Е.Г. Максимова, О.Ю. Шабельникова. // Сибирский медицинский журнал. – 2011. – № 4, Т. 26, Вып. 2. – С. 103–106.

36. Новые подходы в прогнозировании исхода острого панкреатита / Б.Б. Бромберг, Д.Е. Бессонов, Д.С. Криволапов, А.М. Гулько // Бюллетень медицинских Интернет-конференций. – 2013. – № 8, Т. 3. – С. 1043–1044.

37. Заливская А.И. Новая шкала раннего прогнозирования инфицирования панкреонекроза / А.И. Заливская, А.И. Протасевич // Хирургия. Восточная Европа. – 2012. – № 3. – С. 54– 56.

38. Мартиросян В.В. Использование кластерного анализа в оценке статистически значимых связей между комплексом экзогенных факторов и характеристиками острого нарушения мозгового кровообращения / В.В. Мартиросян, Ю.А. Крупская // Забайкальский медицинский вестник. – 2012. – № 2. – С. 83–91.

39. Шафир М.А. Анализ соответствий: представление метода / Шафир М.А. // Социология: 4М. – 2009. – № 28. – С. 29–44.

40. Meter K. van, Schiltz M.-A., Cibois P., Mounier L. The BMS: A History and French Sociological Perspective. In : M. Greenacre, J. Blasius (Eds.), *Correspondence Analysis in the Social Sciences: Recent Developments and Applications*. San Diego, California, USA: Academic Press, 1994. pp. 128–138.

41. Анализ данных. – [Электронный ресурс] – Режим доступа: http://math.nsc.ru/AP/oteks/Russian/links/AD/index.html (дата обращения: 10.08.2014).

42. Benzécri J.-P. *Histoire et Préhistoire de l'Analyse des Données*. Paris: Dunod, 1982. 159 p.

43. Frawley W., Piatetsky-Shapiro G., Matheus C. Knowledge Discovery in Databases: An Overview. *AI Magazine*, 1992, Vol. 13, No. 3, pp. 57–70.

44. Heikki Mannila. Data Mining: machine learning, statistics, and databases. *SSDBM '96 Proceedings of the Eighth International Conference on Scientific and Statistical Database Management*, 1996. pp. 2–9.

45. Fayhad U., Piatetsky-Shapiro G., Smyth P. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 1996, Fall, pp. 37–54.

46. Методы и модели анализа данных: OLAP и Data Mining / [Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И.]. – СПб: БХВ-Петербург, 2004. – 336 с.

47. Диагностика первичной открытоугольной глаукомы с помощью метода инфракрасной спектрометрии / И.Б. Алексеев, Г.М. Зубарева, С.А. Сильченко, А.В. Алексеев // ГЛАУКОМА. – 2010. – № 4. – С. 19–24.

48. Прогнозирование результатов эндовенозной лазерной облитерации у пациентов разных возрастных групп / Е.В. Шайдаков, В.Л. Булатов, Е.А. Илюхин, И.Н. Сонькин, А.Г. Григорян // Новости хирургии. – 2013. – № 2, Т. 21. – С. 61–68.

49. Прогнозирование тяжести течения гипертонической болезни у больных сахарным диабетом 2-го типа методом «деревьев классификации» / С.Н. Коваль, Е.С. Першина, Т.Г. Старченко, А.В. Арсеньев // Експериментальна і клінічна медицина. – 2013. – № 3 (60). – С. 41–45.

50. Метод прогнозирования эффективности восстановительного лечения на основе дерева решений / А.А. Зайцев, Е.Ф. Левицкий, И.А. Ходашинский [и др.] // Вопросы курортологии, физиотерапии и лечебной физической культуры. – 2010. – №5. – С. 35–38.

51. Елисеева Л.Н. Классификация больных, страдающих хронической сердечной недостаточностью методом «деревья классификации» / Л.Н. Елисеева, А.А. Халафян, С.Г. Сафонова // Успехи современного естествознания. – 2006. – № 11. – С. 16–18.

52. Носулич М.С. Применение логических алгоритмов для извлечения знаний из медицинских баз данных / М.С. Носулич, С.Н. Балака, М.Н. Нессонова // «Актуальні питання створення нових ЛЗ» : всеукр. наук.-практ. конф. студентів та молодих вчених; 21 квітня 2011 р., Харків : тези доповідей. – Х., НфаУ, 2011. – С. 511.

53. Система интеллектуального анализа данных для прогнозирования результатов хирургического лечения атеросклероза / К.В. Рудаков, К.В. Ворронцов, М.Р. Кузнецов [и др.]. – [Электронный ресурс] – Режим доступа: http://www.e-expo.ru/docs/sem/vc_ras.pdf (дата обращения: 21.08.2014).

54. Дюк В.А. Data Mining – интеллектуальный анализ данных / В.А. Дюк // Сайт Информационных Технологий. – [Электронный

ресурс] – Режим доступа: http://www.inftech.webservis.ru/it/database/datamining/ar2.htm (дата обращения: 01.09.2014).

55. Варшавский П.Р. Метод поиска решений в интеллектуальных системах поддержки принятия решений на основе прецедентов / П.Р. Варшавский, Р.В. Алехин // International Journal “Information Models and Analyses”. – 2013. – Vol. 2, No. 4. – С. 385–392.

56. Дюк В. Data Mining: ученбный курс / В. Дюк, А. Самойленко. – СПб: Питер, 2001. – 312 с.

57. Сычев О.А. Программно-информационная поддержка процессов идентификации состояния органичной системы: На примере ревматических заболеваний : автореф. дисс. на соискание уч. степени канд. техн. наук : спец. 05.13.01 “Системный анализ, управление и обработка информации (по отраслям)” / Сычев Олег Александрович ; Волгоградский гос. техн. ун-т. – Волгоград, 2005. – 23 с.

58. Александров А.В. Клинико-патогенетическое значение исследования аутоантител к ферментам антиоксидантной системы и пуринового метаболизма у больных ревматоидным артритом : автореф. дисс. на соискание уч. степени д. мед. наук : спец. 14.00.39 “Ревматология” / Александров Андрей Вячеславович ; ГОУ ВПО «Волгоградский государственный медицинский университет» Росздрава. – Волгоград, 2009. – 64 с.

59. Круглов В.В. Искусственные нейронные сети. Теория и практика. – 2-е изд. / В.В. Круглов, В.В. Борисов. – М.: Горячая линия—Телеком, 2002. – 382 с.

60. Хайкин С. Нейронные сети. Полный курс. 2-е изд. / Хайкин С. – М.: Вильямс, 2006. – 1104 с.

61. Постарнак Д.В. Критический анализ моделей нейронных сетей / Постарнак Д.В. // Вестник Тюменского государственного

университета. Физико-математические науки. Информатика. – 2012. – № 4. – С. 162–167.

62. Прогнозная диагностика транзиторных ишемических атак: лечебно-профилактическое предупреждение / Б.В. Дривотинов, Е.Н. Апанель, А.С. Мастыкин, В.А. Головко, Г.Ю. Войцехович // Медицинский журнал. – 2014. – № 1 – С. 9–15.

63. Применение искусственных нейронных сетей для прогнозирования в хирургии / С.Д. Богомолов, С.В. Киселев, А.П. Медведев, В.М. Назаров // Нижегородский медицинский журнал. – 2003. – № 1. – С. 101–107.

64. Беребин М.А. Опыт применения искусственных нейронных сетей для целей дифференциальной диагностики и прогноза нарушений психической адаптации / М.А. Беребин, С.В. Пашков // Вестник ЮУрГУ серия «Компьютерные технологии, управление, радиоэлектроника». – 2006. – № 14, Вып. 4. – С. 41–45.

65. Возможности прогнозирования инфицированного панкреонекроза / Литвин А.А., Жариков О.Г., Сенчук Г.А. [и др.] // Проблемы здоровья и экологии. – 2007. – №2 (12). – С. 7–14.

66. Жариков О.Г. Экспертные системы в медицине / О.Г. Жариков, А.А. Литвин, В.А. Ковалёв // Медицинские новости. – 2008. – №10. – С. 15–18.

67. Шевченко Ю.В. Прогнозирование течения раннего послеоперационного периода у больных с радикальными операциями по поводу рака лёгкого с использованием методов бинарной логистической регрессии и искусственных неейронных сетей : автореф. дисс. на соискание уч. степени канд. мед. наук : спец. 05.13.01 “Системный анализ, управление и обработка информации (по отраслям)” / Шевченко Юрий Владимирович ; ГОУ ВПО «Российский Государственный Медицинский Университет» Федерального агентства по здравоохранению и социальному развитию; Федеральное гос.

учреждение «Российский Научный Центр Рентгенорадиологии» Федерального агентства по высокотехнологичной медицинской помощи. – Москва, 2008. – 25 с.

68. Зиновьев А.Ю. Визуализация многомерных данных / Зиновьев А.Ю. – Красноярск: Изд-во КГТУ, 2000. – 168 с.

69. Сивакова О.Д. Внебольничная пневмония: клинические особенности, фармакоэпидемиологические и фармакоэкономические аспекты в Самарской области: автореф. дисс. на соискание уч. степени канд. мед. наук : спец. 14.01.25 “Пульмонология”; 14.03.06 “Фармакология, клиническая фармакология” / Сивакова Ольга Дмитриевна ; Самарский гос. мед. ун-т. – Самара, 2014. – 24 с.

70. Анализ соответствия и оценка связи эзофагогастродуоденальных заболеваний с наследственными нарушениями соединительной ткани / А.С. Рудой, С.С. Горохов, Д.В. Лапицкий, И.П. Реуцкий // Военная медицина. – 2010. – № 4. – С.59–62.

71. Матуа С.П. Перспективы применения метода многомерного шкалирования при осуществлении электроэнцефалографического мониторинга психофармакотерапии / С.П. Матуа, М.В. Рудковский // Биомедицина. – 2006. – № 4, Т. 1. – С. 95 – 97.

72. Рудковский М.В. Дискретный ЭЭГ мониторинг фармакотерапии психоневрологических больных с использованием метода многомерного шкалирования / М.В. Рудковский, В.П. Омельченко, С.П. Матуа // Изестия ВУЗов Северо-Кавказского региона. Естественные науки. Приложение. – 2003. – № 8.– С. 59–67.

73. Severeyn E., Altuve M., Ng F., et al. Towards the Prediction of Mortality in Intensive Care Units Patients: A Simple Correspondence Analysis Approach. *Proceedings of the CinC (Computing in Cardiology) Conference*, 2012, 9–12 September, Vol. 39, pp. 469–472.

74. Annunziato R.A., Kim S.-K., Fussner M., et al. Utilizing correspondence analysis to characterize the mental health of cardiac patients with diabetes. *Journal of Health Psychology*, 2013, (11), pp. 55–63.

75. Воронцов К.В. Комбинаторная теория надёжности обучения по прецедентам : дисс. … д-ра физ.-мат. Наук : 05.13.17 "Теоретические основы информатики" / Воронцов Константин Вячеславович ; Вычислительный центр РАН. – Москва, 2010. — 271 с.

76. Gordon A.D. *Classification: 2nd ed.* London: Chapman & Hall / CRC Press LLC, 2002. 248 p.

77. Rencher A.C. *Methods of multivariate analysis: 2nd ed.* John Wiley & Sons, Inc. "A Wiley-Interscience publications", 2002. 732 p.

78. Халафян А.А. STATISTICA 6. Статистический анализ данных: 3-е изд. Учебник / Халафян А.А. – М.: ООО «Бином-Пресс», 2007. – 512 с.

79. Орлов А.И. Нечисловая статистика / Орлов А.И. – М.: МЗ-Пресс, 2004. – 513 с.

80. Терехина А.Ю. Анализ данных методами многомерного шкалирования / Терехина А.Ю. – М.: Наука, 1986. – 168 с.

81. Перекрест В.Т. Нелинейный типологический анализ социально-экономической информации: Математические и вычислительные методы / Перекрест В.Т. – Л.: Наука, 1983. – 176 с.

82. Анализ нечисловой информации / [Ю.Н. Тюрин, Б.Г. Литвак, А.И. Орлов, Г.А. Сатаров, Д.С. Шмерлинг]. – М.: Научный Совет АН СССР по комплексной проблеме «Кибернетика», 1981. – 80 с.

83. Borg I., Groenen P.J.F. Modern Multidimensional Scaling: theory and applications, 2nd ed. New York: Springer-Verlag, 2005, 614 p.

84. Torgerson W.S. Multidimensional scaling: A Theory and method. *Psychometrika*, 1952, Vol. 17, No. 3, pp. 401–419.

85. Shepard R.N. The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika*, 1962, Vol. 27, pp. 125–140; pp. 219–246.

86. Многомерный статистический анализ в экономике: Учеб. пособие для вузов / [Сошникова Л.А., Тамашевич В.Н., Уебе Г., Шефер М.] ; под ред. проф. В.Н. Тамашевича. – М.: ЮНИТИ-ДАНА, 1999. – 598 с.

87. Orlóci L. Geometric models in ecology. I. The theory and application of some ordination methods. *J. Ecology*, 1966, No. 54, pp. 193–215.

88. Torgerson W.S. *Theory and methods of scaling*. Michigan U.M.I.: Ann Arbor, 1990. 460 p.

89. Kruskal J.B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 1964, Vol. 29, No. 2, pp. 115–129.

90. Kruskal J.B. Multidimensional scaling by optimizing goodness of fit to a nonparametric hypothesis. *Psychometrika*, 1964, Vol. 29, No. 1, pp. 1–27.

91. Kruskal J. B., Wish M. Multidimensional Scaling. *Sage University Paper series on Quantitative Application in the Social Sciences*, 1978, pp. 7–11.

92. Guttman L. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, 1968, Vol. 33, No. 4, pp. 469–506.

93. Кобзарь А.И. Прикладная математическая статистика [2-е изд., испр.] / Кобзарь А.И. – М.: ФИЗМАТЛИТ, 2012. – 816 с.

94. Трошин Л.И. Статистический анализ нечисловой информации / Трошин Л.И., Балаш В.А., Балаш О.С. – М.: Московский государственный университет экономики, статистики и информатики, 2001. – 67 с.

95. Кендалл М.Дж. Статистические выводы и связи / М.Дж. Кендалл, А. Стюарт. – М.: Наука, Гл. ред. физ.-мат. лит., 1973. – 900 с.

96. The Mathematics of Correspondence Analysis. – [Electronic resource] – Access mode: https://openaccess.leidenuniv.nl/bitstream/handle/1887/17697/ appendix_ references.pdf?sequence=5 (дата обращения: 27.05.2014).

97. Greenacre M. *Correspondence Analysis in Practice, 2nd edition.* London: Chapman & Hall/CRC, 2007. 274 p.

98. Abdi H., Béra M. Correspondence Analysis. In R. Alhajj, J. Rokne (Eds.), *Encyclopedia of Social Networks and Mining.* New York: Springer Verlag, 2014. pp. 275–284.

99. Burt C. The Factorial Analysis of Qualitative Data. *British Journal of Statistical Psychology*, 1950, Vol. 3, No. 3, pp. 166–185.

100. Cattell R.B. The scree test for the number of factors. *Multivariate Behavioral Research*, 1966, No. 1 (2), pp. 245–276.

101. Kaiser H.F. An index of factorial simplicity. *Psychometrika*, 1974, Vol. 39, 31–36.

102. Многомерный статистический анализ. – [Электронный ресурс] – Режим доступа: http://www.pitt.edu/~super7/18011-19001/18831.pdf (дата обращения: 06.03.2014).

103. Электронный учебник StatSoft. – [Электронный ресурс] – Режим доступа: http://www.statsoft.ru/home/textbook/modules/stmulsca.html (дата обращения: 11.02.2014).

104. Harris C.W. On factors and factor scores. *Psychometrika*, 1967, No. 32, pp. 363–379.

105. Леонов В.П. Наукометрика статистической парадигмы экспериментальной биомедицины (по материалам публикаций) / Леонов В.П. // Вестник Томского государственного университета. Серия “Математика. Кибернетика. Информатика”. – 2002. – №275. – С. 17–24.

106. Леонов В.П. Когда нельзя, но очень хочется, или Ещё раз о критерии Стьюдента / Леонов В.П. – [Электронный ресурс] – Режим доступа: http://www.biometrica.tomsk.ru/student.html (дата обращения: 25.04.2014).

107. Kruskal W.H., Wallis A. Use of ranks in one criterion variance analysis. *Journal of the American Statistical Association*, 1952. Vol. 47, pp. 583–621.

108. Mann H.B.; Whitney D.R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*, 1947, 18 (1), 50–60.

109. Singh N. The ratio test for the parameters exponential distributions. *Common. Stat.-Theor. Meth.*, 1985, Vol. 13, No. 6, pp. 116–119.

110. Wald A., Wolfowitz J. On a test whether two samples are from the same population. *Annals of Mathematical Statistics*, 1940, Vol. 11, No. 2, pp. 147–162.

111. Закс Л. Статистическое оценивание / пер. с нем. В.Н. Варыгина; под ред. Ю.П. Адлера, В.Г. Горского. – М.: Статистика, 1976. – 598 с.

112. Холлендер М. Непараметрические методы статистики / М. Холлендер, Д. Вулф. ; пер. с англ. Д.С. Шмерлинга ; под ред. Ю.П. Адлера, Ю.Н. Тюрина. – М.: Финансы и статистика, 1983. – 518 с.

113. Справочник по прикладной статистике / под ред. Э. Ллойда, У. Ледермана ; пер. с англ. под ред. С.А. Айвазяна, Ю.Н. Тюрина. Т. 2. – М.: Финансы и статистика, 1990. – 526 с.

114. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики / Л.Н. Большев, Н.В. Смирнов. – М.: Наука, Гл. ред. физ-мат. лит., 1983 – 416 с.

115. Bonferroni C.E. Teoria statistica del le classi e calcolo del le probabilit`a. *Pubblicazioni del R Istituto Superioredi Scienze Economiche e Commercialidi Firenze*, 1936, Vol. 8, pp. 3–62.

116. Руниoн Р. Справочник по непараметрической статистике: Современный подход / Руниoн Р. ; [пер. с англ. Е.З. Демиденок; предисл. Ю.Н. Тюрина]. – М.: Финансы и статистика, 1982. – 198 с.

117. Аптон Г. Анализ таблиц сопряжённости / Аптон Г. ; [пер. с англ. и предисл. Ю.П. Адлера]. – М.: Финансы и статистика, 1982. – 143 с.

118. Журавлёв Ю.И., Гуревич И.Б. Распознавание. Классификация. Прогноз. Математические методы и их применение. Вып. 2 / Ю.И. Журавлёв, И.Б. Гуревич ; под. ред. Ю.И. Журавлёва. – М.: Наука, 1989. – 69 с.

119. Журавлёв Ю.И. Алгоритмы распознавания, основанные на вычислении оценок / Ю.И. Журавлёв, В.В. Никифоров // Кибернетика. – 1971. – №3. – С. 1–11.

120. Исраилов И.М. Алгоритмы вычисления оценок со сложными системами опорных множеств и их замыкания : дисс. на соискание уч. степени канд. физ.-мат. наук : спец. 01.01.09 “Математическая кибернетика” / Исраилов Илхом Мирхаликович ; Вычислительный центр АН СССР. – Москва, 1985. – 93 с.

121. Нефёдов А.В. Эффективные алгоритмы, основанные на вычислении оценок, с прямоугольными опорными множествами, для задач распознавания изображений : дисс. на соискание уч. степени канд. физ.-мат. наук : спец. 01.01.09 “Дискретная математика и математическая кибернетика” / Нефёдов Алексей Валентинович ; Московский гос. ун-т им. М.В. Ломоносова. – Москва, 2005. – 132 с.

122. Романов М.Ю. Построение обобщённых полиномов минимальной степени над алгоритмами вычисления оценок: автореф. дисс. на соискание уч. степени канд. физ.-мат. наук : спец. 05.13.17 “Теоретические основы информатики” / Романов Михаил Юрьевич ; Вычислительный центр им. А.А. Дородницына РАН. – Москва, 2008. – 19 с.

123. Duda R.O., Hart P.E., Stork D.G. *Pattern Classification*, *2nd ed*. NY John Wiley & Sons, NY, 2001. 680 p.

124. Гуревич И.Б. Схема синтеза логических моделей изображений, допускаемых эффективными распознающими операторами / Гуревич И.Б. // Компьютерная оптика. – 1995. – Вып. 14–15, часть 1. – С. 133–147.

125. Дьяконов А.Г. Алгебра над алгоритмами вычисления оценок: Учеб. Пособие / Дьяконов А.Г. – М.: Изд. отдел ф-та Вычислительной математики и кибернетики МГУ, 2005. – 71 с.

126. Дмитриев А.Н. О математических принципах классификации предметов и явлений / А.Н. Дмитриев, Ю.И. Журавлёв, Ф.П. Кренделев // Дискретный анализ: Сб. статей, Новосибирск: Институт математики СО АН СССР. – 1966. Вып. 7. – С. 3–15.

127. Воронцов К.В. Лекции по алгоритмическим композициям / К.В. Воронцов. – [Электронный ресурс] – Режим доступа: http://www.machinelearning.ru/ wiki/images/0/0d/Voron-ML-Compositions.pdf (дата обращения: 07.10.2012).

128. Родзин С.И. Искусственный интеллект / Родзин С.И. – Таганрог: Изд-во ТТИ ЮФУ, 2009. – 200 с.

129. Ho T.K. Multiple classifier combination: Lessons and the next steps. In : A. Kandel, H. Bunke (Eds.), *Hybrid Methods in Pattern Recognition*. World Scientific Publishing, 2002. pp. 171–198.

130. Tresp V. Committee machines. In : Y.H. Hu, J.-N. Hwang (Eds.), *Handbook for Neural Network Signal Processing*. CRC Press, 2002. pp. 134–151.

131. Мазуров В.Д. Метод комитетов в задачах оптимизации и классификации / Мазуров В.Д. //М.: Наука, 1990, 250 с.

132. Мазуров В.Д. Комитеты системы неравенств и задача распознавания / Мазуров В.Д. // Кибернетика. – 1971. – № 3. – С. 140–146.

133. Osborne M.L. The seniority logic: A logic for a committee machine. *IEEE Trans. on Comp.*, 1977, Vol. C-26, No. 12, pp. 1302–1306.

134. Dietterich T.G. Ensemble methods in machine learning. In : J. Kittler, F. Roli (Eds.), *Multiple Classifier Systems*, vol. 1857 of *Lecture Notes in Computer Science*. Cagliari, Italy: Springer, 2000. pp. 1–15.

135. Chawla N.V., Hall L.O., Bowyer K.W., Kegelmeyer W.Ph. Learning ensembles from bites: A scalable and accurate approach. *Journal of Machine Learning Research*, 2004, No. 5, pp. 421–451.

136. Терехов С.А. Гениальные комитеты умных машин: лекции по нейроинформатике. Часть 2. / Терехов С.А. // Научная сессия МИФИ–2007. IX Всероссийская научно-техническая конференция «Нейроинформатика–2007». – М.: МИФИ, 2007. – 148 с.

137. Kittler J., Roli F. (Eds.) Multiple Classifier Systems. *Proceedings of the Second International Workshop, MCS 2001*, Cambridge, UK, July 2–4, 2001. 454 p.

138. Jain A.K., Duin R.P.W., Mao J. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, Vol. 22, No. 1, pp. 4–37.

139. Барабаш Ю.Л. Коллективные статистические решения при распознавании / Барабаш Ю.Л. – М.: Радио и связь, 1983. – 224 с.

140. Battiti R., Colla A.M. Democracy in neural nets: Voting schemes for classification. *Neural Networks*, 1994, No. 7, pp. 691–707.

141. Lam L., Krzyzak A. A theoretical analysis of the application of majority voting to pattern recognition. *Proceedings of the 12th*

International Conference on Pattern Recognition. Jerusalem, Israel, 1994. pp. 418–420.

142. Lam L., Suen C.Y. Application of majority voting to pattern recognition: An analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics*, 1997, No. 27, Vol. 5, pp. 553– 568.

143. Lin X., Yacoub S., Burns J., Simske S. Performance analysis of pattern classifier combination by plurality voting. *Pattern Recognition Letters*, 2003, No. 24, Vol. 12, pp. 1795–1969.

144. Ruta D., Gabrys B. A theoretical analysis of the limits of majority voting errors for multiple classifier systems. *Pattern Analysis & Applications*, 2002, Vol. 5 (Issue 4), pp. 333–350.

145. Мазуров В.Д. Комитетные конструкции / В.Д. Мазуров, М.Ю. Хачай // Известия УрГУ. – 1999. – № 14. – С. 77–108.

146. Растригин Л.А. Коллективные правила распознавания / Л.А. Растригин, Р.Х. Эренштейн. – М.: Энергия, 1981. – 244 с.

147. Littlestone N., Warmuth M.K. The weighted majority algorithm. *IEEE Symposium on Foundations of Computer Science*. 1989. pp. 256–261.

148. Гончаров М. Ансамбли моделей / Гончаров М. – [Электронный ресурс] – Режим доступа: http://businessdataanalytics.ru/download/ModelEnsembles.pdf (дата обращения: 23.12.2013).

149. Rivest R.L. Learning decision lists. *Machine Learning*, 1987, Vol. 2, No. 3, pp. 229–246.

150. Marchand M., Shawe-Taylor J. Learning with the set covering machine. *Proceedings of the 18th International Conference on Machine Learning*. San Francisco, CA, 2001. pp. 345–352.

151. Комарцова Л.Г. Построение системы обнаружения вторжений в локальную сеть на основе нейросетевых ассоциативных машин /

Л.Г. Комарцова, Ю.Н. Лавренков // Нейроинформатика. – 2013. – Ч. 3. – С. 129–139.

152. Jacobs R.A., Jordan M.I., Nowlan S.J., Hinton G. E. Adaptive mixtures of local experts. *Neural Computation*, 1991,No. 3, pp. 79–87.

153. Gamma J., Brazdil P. Cascade generalization. *Machine Learning*, 2000, No. 41(3), pp. 315–343.

154. Jordan M.I.,Xu L. Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 1995, No. 8, pp. 1409–1431.

155. Nowlan S.J., Hinton G.E. Evaluation of adaptive mixtures of competing experts. In : R.P. Lippmann, J.E. Moody, D.S. Touretzky (Eds.), *Advances in Neural Information Processing Systems*, 1991. pp. 774–780.

156. Аркадьев А.Г. Обучение машины классификации объектов / А.Г. Аркадьев, Э.М. Браверман. – М.: Наука, Гл. ред. физ.-мат. лит., 1971. –192 с.

157. Аркадьев А.Г. Обучение машины распознаванию образов / А.Г. Аркадьев, Э.М. Браверман. – М.: Наука, 1964. – 110 с.

158. Воронцов К.В. Лекции по метрическим алгоритмам классификации / К.В. Воронцов. – [Электронный ресурс] – Режим доступа: http://www.machinelearning.ru/wiki/images/c/c3/Voron-ML-Metric-slides.pdf (дата обращения: 19.12.2014).

159. Ильченко А.В. Классификация на основе компонентных структур данных в признаковом пространстве / Ильченко А.В. // Таврический Вестник Информатики и Математики. – 2004. – № 2. – С. 163–171.

160. Лбов Г.С. Группировка объектов в пространстве разнотипных признаков / Г.С. Лбов, Т.М. Пестунова [в кн. Анализ нечисловой

информации в социологических исследованиях]. – М.: Наука, 1985. – 226 с.

161. Устенко А.С. Основы математического моделирования и алгоритмизации процессов функционирования сложных систем / Устенко А.С. – М.: Изд-во “МИР”, 2000. – 555 с.

162. Мурыгин К.В. Построение классификаторов на основе разделяющих поверхностей / Мурыгин К.В. // Штучний інтелект. – 2008. – № 2. – С. 65–69.

163. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию / М. Шлезингер, В. Главач. – К.: Наукова думка, 2004. – 554 с.

164. Местецкий Л.М. Математические методы распознавания образов: Курс лекций. / Местецкий Л.М. // МГУ, ВМиК, кафедра «Математические методы прогнозирования», 2002–2004. – [Электронный ресурс] – Режим доступа: http://www.ccas.ru/frc/papers/mestetskii04course.pdf (дата обращения: 24.02.2014).

165. Cover T., Hart P. Nearest neighbor pattern classification. *Information Theory, IEEE Tran*, 1967, Vol. 13 (Issue 1), pp. 21–27.

166. Bremner D., Demaine E., Erickson J., et al. Output-Sensitive Algorithms for Computing Nearest-Neighbour Decision Boundaries. *Discrete & Computational Geometry*, 2005,Vol. 33 (Issue 4), pp. 593–604.

167. Dasarathy B.V. [ed.]. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, 1991. 447 p.

168. Shakhnarovish G., Darrell T., Indyk P. [ed.] *Nearest-Neighbor Methods in Learning and Vision*. Cambridge, Massachusetts London, England: MIT Press, 2006. 280 p.

169. Beyer K., Goldstein J., Ramakrishnan R., Shaft U. When Is "Nearest Neighbor" Meaningful? *Proceedings from Database Theory – ICDT'99 : 7th International Conference,* Jerusalem, Israel, 1999, January 10–12, pp. 217–235.

170. Hastie T.,Simard P.Y. Metrics and Models for Handwritten Character Recognition. *Statistical Science*, 1998, No. 13, pp. 54–65.

171. Волошин Г.Я. Методы распознавания образов: Конспект лекций / Волошин Г.Я. // Сервер Методического Обеспечения ВГУЭС. – [Электронный ресурс] – Режим доступа: http://abc.vvsu.ru/Books/Metody_r/ (дата обращения: 14.03.2014).

172. Загоруйко Н.Г. Прикладные методы анализа данных и знаний / Загоруйко Н.Г. – Новосибирск: ИМ СО РАН, 1999. – 264 с.

173. Vapnik V.N. An Overview of Statistical Learning Theory. *IEEE Transactions On Neural Networks*, 1999, Vol. 10, No. 5, pp. 988–999.

174. Bellman R.E. *Adaptive Control Processes: A Guided Tour.* Princeton, NJ: Princeton University Press, 1961. 274 p.

175. *Дэйвисон М. Многомерное шкалирование: методы наглядного представления данных / Дэйвисон М. – М. : Финансы и статистика, 1988. – 254 с.*

176. Толстова Ю.Н. Основы многомерного шкалирования / Толстова Ю.Н. – М. : КДУ, 2006. – 160 с.

177. Сатаров Г.А. Многомерное шкалирование и другие методы при комплексном анализе данных / Сатаров Г.А. [в кн. Анализ нечисловой информации в социологических исследованиях]. – М.: Наука, 1985. – 226 с.

178. Greenacre M. *Theory and Applications of Correspondence Analysis*. London: Academic Press, 1984. 364 p.

179. Фёрстер Э. Методы корреляционного и регрессионного анализа. / Э. Фёрстер, Б. Рёнц. – М.: Финансы и статистика, 1983. – 303 с.

180. Дрейпер Н. Прикладной регрессионный анализ. – 3-е изд. / Н. Дрейпер, Г. Смит. – М.: Диалектика, 2016. –912 с.

181. Brown T.A. *Confirmatory Factor Analysis for Applied Research*. New York: The Guilford Press, Inc., 2006. 493 p.

182. Афифи А. Статистический анализ: Подход с использованием ЭВМ. / А. Афифи, С. Эйзен. – М.: Мир, 1982. – 488 с.

183. Дубров А.М. Многомерные статистические методы: Учебник. / А.М. Дубров , В.С. Мхитарян, Л.И. Трошин. – М. : Финансы и статистика, 2003. – 352 с.

184. Журавлёв Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. / Ю.И. Журавлёв, В.В. Рязанов, О.В. Сенько. – М. : Изд. ФАЗИС, 2005. – 159 с.

185. Колмогоров А.Н Элементы теории функций и функционального анализа. / А.Н. Колмогоров, С.В. Фомин – М. : Наука, 1976. – 543 с.

186. Folland G.B. *Real Analysis: Modern Techniques and Their Applications, 2nd ed.* John Wiley & Sons, Inc., 1999. – 408 p.

187. Cormen T.H., Leiserson Ch.E., Rivest R.L., Stein C. *Introduction to Algorithms, 3rd Edition*. Cambridge, Massachusetts London, England: MIT Press, 2009. 1292 p.

188. Рутковская Д. Нейронные сети, генетические алгоритмы и нечеткие системы: Пер. с польского И.Д. Рудинского / Д. Рутковская, М. Пилиньский, Л. Рутковский – М.: Горячая линия—Телеком, 2004. – 452 с.

189. Перепелица В.А. Дискретная оптимизация и моделирование в условиях неопределенности данных / В.А. Перепелица, Ф.Б. Тебуева. – М.: Изд-во «Академия Естествознания», 2007 – 151 с.

190. *Алтунин А.Е., Семухин М.В.* Модели и алгоритмы принятия решений в нечётких условиях / *А.Е. Алтунин, М.В. Семухин* – Тюмень: Изд-во ТюмГУ, 2000. – 352 с.

191. Titterington D.M., Murray G.D., Murray L.S., Spiegelhalter D.J., Skene A.M., Habbema J.D.F., Gelpke G.J. Comparison of discriminant techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society, Series A (General)*, 1981, Vol. 144(No. 2), pp. 145–175.

192. Domingos P., Pazzani M. On the optimality of the simple Bayesian classifier under zero–one loss. *Machine Learning*, 1997, 29, pp. 103–130.

193. Ripley B.D. *Pattern Recognition and Neural Networks*. Cambridge: University Press, 2007. 403 p.

194. Конспект лекций по курсу «Методы и средства анализа данных» / кафедра Информационно-коммуникационных технологий Московского государственного института электроники и математики. – [Электронный ресурс] – Режим доступа: http://wiki.auditory.ru/ Конспект_лекций_ по_курсу_“Методы_и_средства_анализа_данных” (дата обращения: 03.02.2014).

195. Дюк В.А. Осколки знаний / Дюк В.А. // Экспресс-Электроника. – 2002. – № 6. – С. 60–65.

196. Стасышин В.М. Методы построения деревьев решений в задачах классификации в Data Mining / Стасышин В.М. // методические материалы кафедры программных систем и баз данных Новосибирского государственного технического университета. – [Электронный ресурс] – Режим доступа: http://www.ami.nstu.ru/~vms/lecture/data_mining/trees.htm (дата обращения: 30.01.2014).

197. Воронцов К.В. Лекции по логическим алгоритмам классификации / К.В. Воронцов. – [Электронный ресурс] – Режим

доступа: http://www.ccas.ru/voron/download/LogicAlgs.pdf (дата обращения: 01.12.2017).

ПЕРЕЧЕНЬ ИСПОЛЬЗВАННЫХ УСЛОВНЫХ СОКРАЩЕНИЙ

CART – полный перебор одномерных ветвлений при построении дерева классификации (Classification And Regression Trees)

FACT – метод прямой остановки ветвления при построении дерева классификации

GCS – шкала комы Глазго (Glasgow Coma Score)

KDD – методы обнаружения знаний в базах данных (Knowledge Discovery in Databases)

NIHSS – National Institutes of Health Stroke Scale (шкала тяжести инсульта)

SIRS – Systemic Inflammatory Response Syndrome (шкала критериев тяжести сепсиса)

SOFA – Sepsis-related Organ Failure (шкала оценки полиорганной недостаточности)

АВО – алгоритм вычисления оценок

АД – артериальное давление

АЛТ – аланиновая трансаминаза крови

ДА – дискриминантный анализ

КА – корреспондентский анализ

МИС – медицинская информационная система

МШ – многомерное шкалирование

ПЖ – поджелудочная железа

ПО – программное обеспечение

ТПЖ – травма поджелудочной железы