DOI: 10.18720/SPBPU/2/id20-61

1. APPLICATION OF MACHINE LEARNING METHODS IN MEDICINE

***Oleg Valentinovich Senko**, dr. sci. (phys.–math.), leading researcher, Federal Research Center "Informatics and Control" of Russian Academy of Sciences, Russia, Moscow, Vavilova st. 44-2, BOX 119333, senkoov@mail.ru.*

***Anna Victorovna Kuznetsova**, cand. sci. (biology), senior researcher, N.M. Emanuel Institute of Biochemical Physics RAS, Russia, Moscow, Kosygina st. 4, BOX 119334, azfor@yandex.ru.*

Annotation. *The chapter discusses the applications using machine learning technologies in various tasks of medical diagnostics and forecasting. Machine learning methods provide computer-made construction of algorithms predicting unknown target indicator by known values of another variables. Such algorithms generated from empirical regularities discovered in data. The chapter presents a short review of existing mainstream technologies including statistical methods, artificial neural networks, classification trees and forests, Bayesian networks. Original technologies are based on collective solutions by sets of regularities. Questions of correct evaluation of diagnostic or predicting algorithms efficiency are discussed. Also the paper discusses existing methods of informative variables search.*

Keywords. *Machine learning, prediction, data base, diagnostics algorithm.*

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В МЕДИЦИНЕ

***Олег Валентинович Сенько**, д.ф.-м.н., ведущий научный сотрудник, ФИЦ "Информатика и управление" Российской академии наук, Россия, Москва, Вавилова 44 к.2, индекс 119333, senkoov@mail.ru.*

***Анна Викторовна Кузнецова**, к.б.н., старший научный сотрудник, Институт биохимической физики им. Н.М. Эмануэля Российской академии наук, Россия, Москва, ул. Косыгина, д. 4, индекс 119334, azfor@yandex.ru.*

Аннотация. *В главе рассматривается использование технологий машинного обучения, предсказывающих неизвестные значения целевого показателя по известным значениям других показателей. Генерация таких алгоритмов производится на основании эмпирических закономерностей, обнаруживаемых в данных. В главе приводится краткий обзор наиболее распространённых технологий машинного обучения, включая статистические методы, искусственные нейронные сети, метод опорных векторов, решающие деревья и леса, байесовские сети. Рассматриваются также оригинальные технологии, основанные на принятии коллективных решений по системам закономерностей. Обсуждаются вопросы корректной оценки эффективности алгоритмов диагностики и прогноза. Приводятся современные методы отбора наиболее информативных показателей.*

Ключевые слова. *Машинное обучение, прогнозирование, базы данных, диагностические алгоритмы.*

Введение

В главе приводится краткий обзор наиболее распространённых технологий машинного обучения (далее — МО), включая статистические методы, искусственные нейронные сети, метод опорных векторов, решающие деревья и леса, байесовские сети. Рассматриваются также оригинальные технологии, основанные на принятии коллективных решений по системам закономерностей. Обсуждаются вопросы корректной оценки эффективности алгоритмов диагностики и прогноза. Приводятся современные методы отбора наиболее информативных показателей.

Создание медицинских баз данных, включающих значения показателей различной природы, поступающих из многих источников, вызывает необходимость использования математических и программных средств, позволяющих извлечь из собранной нередко весьма разнородной информации максимальную пользу. Одним из способов использования медицинских баз данных с целью повышения эффективности лечения является применение методов МО. МО является областью искусственного интеллекта, занимающейся построением математических моделей на основании имеющихся данных.

Целью использования МО в медицине является, в основном, построение по накопленным клиническим данным моделей и алгоритмов для осуществления компьютерной диагностики или прогноза по доступной объективной информации о пациенте. Построение и настройка моделей и алгоритмов производится на основе существующих в данных эмпирических закономерностей, как правило, в автоматическом режиме без непосредственного участия человека

Термин *машинное обучение* возник при построении компьютерных алгоритмов с дальнейшей их настройкой по аналогии с обучением живых организмов в процессе их адаптации в природе. Впервые термин был введён Артуром Самуэлем (Arthur Samuel) в 1959 году применительно к задаче автономной настройки компьютерной программы для игры в шашки [1.25]. Примерно в это же время Ф. Розенблаттом была предложена математическая модель нейрона [1.23], компьютерные реализации которой далее достаточно успешно использовались в задачах диагностики и прогнозирования в самых различных областях. Становление методов машинного обучения как научной дисциплины происходило во второй половине прошлого века, когда оформились основные направления, включая статистические модели, основанные на байесовском обучении, нейросетевые методы, решающие деревья и леса, метод опорных векторов, комбинаторно-логические модели и др. Интенсивное развитие указанных подходов происходит в настоящее время.

Существующие в настоящее время средства машинного обучения позволяют решать разнообразные задачи медицинской диагностики и прогнозирования по весьма разнородной информации, включая клинические, биохимические, генетические показатели, электрические сигналы организма; данные ультразвуковой диагностики, рентгеновские или гистологические изображения, показатели, характеризующих образ жизни пациентов, а также показатели, описывающие использованные схемы лечения и др. В число таких задач входят задачи прогнозирования: риска возникновения заболеваний, результатов лечения и исхода заболевания, задачи компьютерной диагностики заболеваний, включая дифференциальную диагностику.

При решении задач прогнозирования результатов лечения анализируют прогнозы, соответствующие различным схемам, что позволяет подобрать оп-

тимальную для каждого случая стратегию лечения. Решение задач прогнозирования возникновения заболеваний позволяет оценить основные риски для конкретного пациента. Таким образом, алгоритмы компьютерного прогнозирования, полученные с помощью методов МО, являются важным средством поддержки принятия решений при назначении индивидуальных курсов лечения или организации целенаправленной профилактики. Накопленная за многие годы клинико-лабораторная информация, таким образом, перестает быть просто архивными данными, а начинает активно работать на повышение эффективности лечебного процесса.

Использование методов МО потенциально позволяет существенно увеличить точность диагностики и качество лечения за счёт учёта индивидуальных особенностей каждого пациента [1.7]. Наличие технологий прогнозирования исхода заболевания позволяет объективно оценить тяжесть заболевания, что в дальнейшем дает возможность сравнивать работу отдельных лечебных учреждений по эффективности с учетом тяжести заболевания проходивших лечение пациентов.

Еще одной областью, где машинное обучение может принести значительную пользу, является создание методов компьютерной диагностики заболевания, включая дифференциальную диагностику по наборам разнообразных клинических, лабораторных и инструментальных показателей. Преимуществом компьютерных средств диагностики является возможность учёта большого количество объективных показателей в их совокупности и взаимосвязях [1.18]. Безусловно, компьютерные методы могут быть только вспомогательным инструментом поддержки принятия решения лечащим врачом. Однако, их использование упрощает процесс диагностики и позволяет избежать ошибочных решений, связанных с неверной оценкой каких-то сочетаний показателей.

Перечисленные задачи принято также рассматривать как задачи классификации или распознавания. Под распознаванием понимается задача отнесения объекта к одной из нескольких категорий (классов). Задача диагностики является задачей отнесения диагностируемого случая к одному из нескольких видов заболевания. Задача прогнозирования результата лечения на самом деле является задачей отнесения диагностируемого случая к одному из

типов результата. Например, результат может характеризоваться как успешный или неуспешный. Необходимым условием использования методов МО является наличие базы данных, содержащих по возможности полные описания случаев, для которых производится прогнозирование. Такая база служит источником для формирования обучающей выборки.

1.1. Требования к данным при использовании методов МО

Данные, вносимые в базу, должны включать значения показателей, по которым вычисляется прогноз, а также значения той величины, которую мы собираемся прогнозировать. Например, при решении задачи прогнозирования исхода лечения для каждого случая в базу данных должны быть включены соответствующие значения предполагаемых прогностических факторов, а также показатель, характеризующий результат, т.е. исход лечения каждого пациента (целевая функция). При этом в базу данных необходимо вносить значения прогнозирующих факторов, известные на момент, когда производится прогноз. Предположим, что создаётся алгоритм прогнозирования исхода лечения по результатам первичного обследования, проведённого на момент поступления пациента в лечебное учреждение. Тогда в базу данных могут вноситься только значения факторов, полученные в ходе первичного обследования.

При решении задачи диагностики наряду с диагностическими показателями должен быть известен точно установленный диагноз. Т.е. в обучающую выборку включают только подтвержденные диагнозы. Для обеспечения корректности использования методов МО данные, собранные в базу, должны, пройти проверку на использование только допустимых значений показателей, на правильное заполнение пропущенных значений, на повторное включение в базу описаний одних и тех же случаев.

Для большинства методов МО непосредственно допустимыми являются числовые показатели. Используемые нечисловые (качественные) показатели должны быть заменены числовыми. Для бинарных показателей, указывающих на наличие или отсутствие какого-либо свойства A, такая замена является тривиальной: достаточно сопоставить 1 случаям, обладающим свой-

ством A, 1 и сопоставить 0 случаям, свойством A не обладающим, 0. В случае, когда качественная переменная может принимать несколько нечисловых значений, задача замены этой переменной оказывается не столь тривиальной. Простая однозначная замена символов на числа может оказаться не оптимальной. Это допустимо только в том случае, когда нечисловые обозначения представляют собой ряд какого-то возрастающего или убывающего по значениям качества. В противном случае возможным вариантом является замена одной многозначной качественной переменной, принимающей значения из набора $A_1, \ldots, A_n$ на n бинарных переменных, указывающих на наличие или отсутствие свойств $A_1, \ldots, A_n$. В этом случае один показатель будет заменен на n производных показателей. Под пропусками понимаются значения показателей, которые по каким-либо причинам не были определены.

Распространённой причиной появления пропущенных значений (пропусков) является, например, отсутствие результатов каких-то лабораторных анализов или инструментальных обследований. Обычно пропуски в базе данных заменяются определенным символом. Многие известные методы МО не могут быть напрямую использованы при обучении по данным с пропущенными значениями, и требуют замены символа пропуска на числовое значение (чаще всего, среднее по показателю). В настоящее время существуют различные методы замены пропусков [1.24], применяемые обычно в автоматическом режиме. Однако для корректной их обработки необходимо, чтобы для обозначения пропуска всегда использовался один и тот же символ, который должен корректно передаваться программе, осуществляющей замену.

Дублирование случаев в базе данных может приводить к существенному завышению оценок точности прогноза при использовании режима кросс-валидации (см. подраздел 1.2), а также к отклонению сгенерированного алгоритма диагностики или прогноза от оптимального. Наконец, существенную пользу при подготовке данных могут принести методы поиска выпадающих наблюдений. Под выпадающим наблюдением понимается значение некоторого показателя (выброс), которое значительно отклоняется от основной доли значений этого показателя в данных. Выпадающие значения часто связаны с ошибками при заполнении базы. В этих случаях целесообразно попытаться восстановить истинные значения показателей. Проверку на выпадаю-

щие наблюдения осуществляют по критерию «трех сигм», критерию Диксона и критерию Граббса [1.4]. На рисунке 1.1 представлен пример базы данных, подготовленный в соответствии с перечисленными выше требованиями с помощью приложения Data Master Azforus (см. подраздел 1.7).

№ объекта	group	3	4	5	6	7	8	9	10	11	12	14	аа_13	аб_13	ас_13
1	1	350	3	-6	3	34	4	333	433	0	2	0	1	0	0
2	1	7	4	-7	4	9,4	17,1	12	43,9	0	2	0	1	0	0
3	1	123	32	-6	12,4	6,6	19	30	29,5	-1000	-1000	0	0	1	0
4	1	213	1	-4	7,7	9,4	17,1	12	43,9	0	-1000	0	0	0	1
5	2	168	4	-8	2	55	5	65	3	0	33	0	0	1	0
6	2	234	1	-4	11,4	12,6	24	65	47,03	0	2	0	0	0	1
7	2	200	1	-2	4	7,2	11,2	20	33,9	0	-1000	0	0	1	0

Рис. 1.1. Вариант базы данных, подготовленной для анализа с использованием методов МО

Целевой переменной в данном случае является переменная «group», представленная в третьем столбце. Переменные $3, 4, \ldots, \text{аб}_{13}, \text{ас}_{13}$ могут быть использованы для прогноза целевого показателя. Для обозначения пропуска используется число -1000. Выбор именно этой замены для пропуска связан с тем, что ни один из показателей не принимает такого значения. Из базы данных, подготовленной в соответствии с изложенными выше требованиями, формируется обучающая выборка, по которой затем производится обучение, то есть построение алгоритма, осуществляющего диагностику или прогноз.

1.2. Методы оценки эффективности алгоритмов, полученных с помощью технологий МО

Оценка результатов работы алгоритма должна производиться на контрольной выборке, содержащей описания только новых случаев прогнозируемого или диагностируемого заболевания, не представленных в обучающей выборке. Включение в контрольную выборку случаев из обучающей выборки обычно приводит к значительному искажению истинной точности прогноза или диагностики в сторону завышения.

Рассмотрим основные, используемые в машинном обучении критерии эффективности диагностики и прогноза, которые должны вычисляться на кон-

трольной выборке. Для оценки эффективности работы может использоваться точность распознавания в смысле доли правильных отнесений в классы. В англоязычной литературе данную характеристику принято обозначать термином accuracy [1.35]. Доля правильных прогнозов хорошо описывает ситуацию, когда классы (категории) примерно одинаково представлены в контрольной выборке. В случае, когда доля объектов одного из классов значительно превышает доли других классов, то использование accuracy может приводить к нелепым результатам.

Рассмотрим задачу распознавания двух классов, один из которых представлен в контрольной выборке значительно чаще, чем другой. Например, доля объектов первого класса в контрольной выборке составляет 91%, а второго только 9%. Индекс accuracy при простом отнесении всех случаев в первый класс, будет равен 91%. Индекс accuracy для действительно высокоточного алгоритма, правильно распознающего 90% объектов каждого из двух классов, будет равен 90% и окажется, таким образом, ниже индекса accuracy, соответствующего отнесению всех объектов в один и тот же класс.

Поэтому, наряду с индексом accuracy используются также другие критерии. Часто при использовании методов МО ставится цель правильно диагностировать какие-либо нежелательные случаи, соответствующие, например, тяжёлому заболеванию или неблагоприятному исходу из совокупности случаев со сходной симптоматикой. Такого рода случаи мы далее будем называть целевыми. Совокупность всех целевых случаев назовём целевым классом. Долю правильно диагностированных целевых случаев принято называть чувствительностью. В англоязычной литературе наряду с термином чувствительность («sensitivity») используется также термин «recall» [1.35]. Долю правильно диагностированных случаев, не принадлежащих целевому классу, принято называть специфичностью.

Наряду с критериями «чувствительность» и «специфичность» в англоязычной литературе используется также термин «precision», обозначающий долю правильно диагностированных случаев, среди всех случаев, классифицированных как целевые. Поясним смысл понятий «accuracy», «sensitivity» и «precision» на примере задачи прогнозирования возникновения новых

осложнений в течение полугода после перенесенного обострения ишемической болезни сердца.

В данной задаче эффективность оценивалась по выборке 1193 пациентов, среди которых у 136 в течение полугода наблюдались осложнения, у 1057 пациентов осложнений за этот период соответственно не было. С использованием метода кросс-валидации (см. ниже) возникновение осложнений было предсказано для 82 пациентов, у которых осложнения произошли, и для 317 пациентов, у которых осложнений не было. Общее число правильных прогнозов составило $82 + 1057 - 317 = 822$, тогда процентная доля правильных прогнозов «accuracy», чувствительность «sensitivity» и точность в смысле «precision» соответственно равны:

- accuracy= $(822 * 100)/1193 = 68{,}9\%$;
- sensitivity= $82/136 = 60\%$;
- precision= $82/(82 + 317) = 20{,}5\%$.

Другим общепринятым критерием эффективности диагностического алгоритма в теории машинного обучения является F-мера, определяемая как среднегармоническое критериев «recall» и «precision», то есть справедливо (1.1).

$$F = 2\frac{\text{recall} * \text{precision}}{\text{recall} + \text{precision}} \tag{1.1}$$

Критерии чувствительности, специфичности или F-меры более верно характеризуют успешность распознавания целевых случаев, чем критерий точности в смысле термина accuracy. Однако полной картины эффективности использования метода машинного обучения указанные критерии также не дают. Дело в том, что каждый алгоритм распознавания выполняется в два этапа.

На первом этапе процесса распознавания вычисляется мера сходства распознаваемого случая с целевым, которую мы обозначим через γ . Величину γ также принято называть «оценкой за целевой класс»[1.31] . На втором этапе происходит окончательная диагностика с помощью решающего правила. Решающее правило состоит в сравнении оценки за целевой класс γ с порогом δ . Распознаваемый случай относится к целевому классу в случае выполнения неравенства $\gamma > \delta$. При невыполнении указанного неравенства распознаваемый случай не считается целевым.

При повышении порога δ общее число случаев, относимых целевому классу, очевидно, сокращается, но возрастает число случаев, не отнесённых к целевым. При этом чувствительность алгоритма распознавания сокращается, а специфичность возрастает. При снижении порога δ наоборот чувствительность алгоритма распознавания возрастает, а специфичность снижается. Вместо критерия специфичность может быть использован критерий «ложная тревога», обозначающий долю случаев, не относящихся к целевым, ошибочно отнесённых целевому классу. Значение критерия «ложная тревога» может быть получено вычитанием значения специфичности из единицы.

Выбор оптимального порогового правила зависит от наших конечных целей. Если нашей целью является по возможности более полное выделение целевого класса, то нам необходимо использовать низкие значения порога δ. Так, низкие значения порога нужно использовать, если мы стремимся выявить по возможности всех пациентов с высоким риском неблагоприятного исхода. Если наша цель – только распознавание случаев с наибольшей вероятностью принадлежащих целевому классу, то нам необходимо использовать высокие значения порога δ .

Перебирая всевозможные пороговые значения, получаем множество пар значений чувствительности и ложной тревоги. Отобразив эти пары на плоскости в соответствующих координатных осях, получим набор точек. Соединив эти точки отрезками прямых, получаем ломаную линию, которую принято называть ROC-кривой [1.8; 1.35]. Название ROC является аббревиатурой английского названия receiver operating characteristic (рабочая характеристика приёмника). Анализ классификаторов через построение ROC-кривых принято называть ROC-анализом. Пример ROC кривой для алгоритма, прогнозирующего возникновение новых осложнений в течение полугода после перенесенного обострения ишемической болезни сердца [1.9], представлен на рисунке 1.2.

Из рисунка 1.2 видно, что 60% всех случаев возникновения осложнений в течение полугода выявляется при уровне ложной тревоги, равном 30%. Видно также, что при уровне ложной тревоги, равном 50%, выявляется свыше 80% случаев возникновения осложнений.

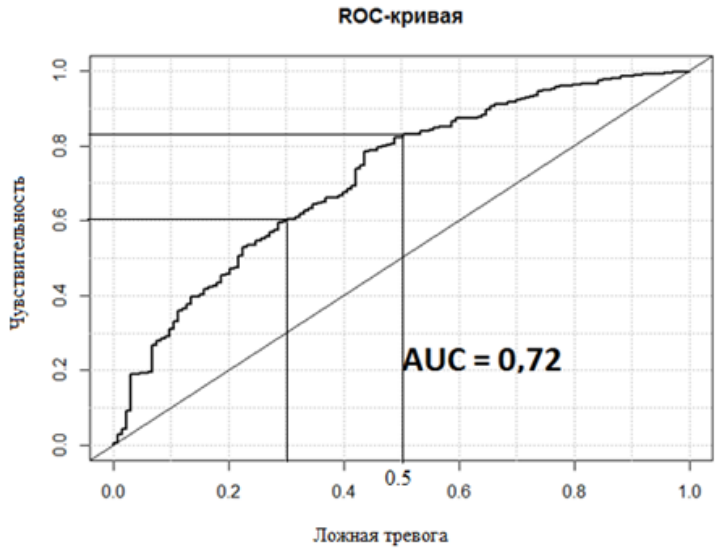

Рис. 1.2. Пример ROC кривой для задачи прогнозирования повторения ОКС в течение полугода после выписки их стационара из работы [1.9]

ROC–кривая позволяет увидеть, при каких уровнях ложной тревоги может достигаться тот или иной уровень чувствительности. Кроме того, ROC-анализ позволяет оценить эффективность использования метода машинного обучения для рассматриваемой задачи в целом, то есть безотносительно к конкретному решающему правилу. Очевидно, что чем выше уровень чувствительности, при каждом заданном уровне ложной тревоги, тем эффективнее используемый метод машинного обучения. В свою очередь более высокое прохождение ROC–кривой соответствует большей площади под ней. Поэтому в качестве меры эффективности того или иного метода машинного обучения для рассматриваемой задачи принято использовать параметр AUC (area under curve), равный площади под ROC–кривой. Для задачи прогнозирования повторного возникновения острого коронарного синдрома (далее — ОКС) в течение полугода параметр AUC равен 0,72. Параметр AUC может быть вычислен по контрольной выборке. Для этого достаточно знать оценки за целевой класс и информацию об истинной принадлежности к целевому классу представленных в контрольной выборке случаев.

Подчеркнём ещё раз, что перечисленные выше параметры «точность» («accuracy»), «чувствительность», «специфичность», AUC отображают истин-

ную точность алгоритма распознавания или метода машинного обучения только в том случае, если они рассчитаны на новых случаях, которые не использовались для настройки (обучения) алгоритмов.

На практике обучающая и контрольная выборки формируются из собранной базы данных с помощью случайного отбора случаев в каждую из них. Размеры этих выборок определяются отдельно для каждой конкретной задачи. Например, можно сформировать равные по размеру обучающую и контрольную выборки. Недостатком такого подхода при ограниченном числе случаев, представленных в исходной базе данных, является снижение соответствующих размеров обучающей и контрольной выборок. Снижение числа случаев в обучающей выборке приводит к снижению точности настройки параметров компьютерной модели диагностики или прогноза. Полученные в результате настройки значения параметров могут существенно отклоняться от оптимальных значений. Снижение размеров контрольной выборки в свою очередь ведёт к увеличению погрешности оценки рассмотренных выше параметров эффективности диагностики или прогноза.

Альтернативным подходом является использование метода k-блоковой кросс-проверки (кросс-валидации). В англоязычной литературе для обозначения k-блоковой кросс-проверки используется термин «k-fold cross validation»» [1.12]. При этом случаи, представленные в исходной выборке случайным образом, разбиваются на k приблизительно равных частей (блоков).

На первом шаге в качестве контрольной выборки используется первый из блоков. Обучающая выборка формируется из всех блоков за исключением первого. Обученный алгоритм производит классификацию случаев из контрольной выборки.

На втором шаге в качестве контрольной выборки используется второй из блоков. Обучающая выборка формируется из всех блоков за исключением второго. Процесс продолжается до тех пор, пока все объекты исходной выборки не окажутся классифицированными. Сравнение результатов распознавания обученными алгоритмами с истинной принадлежностью удаляемых объектов к их классам позволяет вычислить значения параметров «точности», «чувствительности», «специфичности». Использование вычис-

ленных оценок за классы позволяет вычислить ROC-кривую и оценить величину AUC.

1.3. Основные технологии машинного обучения

1.3.1. Статистические методы

Байесовские методы. Статистические методы основаны на оценивании вероятностей принадлежности случая распознаваемым классам при заданных значениях показателей, по которым производится распознавание. Такую вероятность часто называют апостериорной. Классификация производится по простому правилу: диагностируемый случай относится в тот класс, для которого апостериорная вероятность максимальна. Данное правило, которое называют байесовским классификатором [1.12], обеспечивает максимально возможную точность распознавания в смысле термина «accuracy», если подлинные значения апостериорных вероятностей известны. То есть, байесовский классификатор обеспечивает максимальную долю правильных ответов при известной апостериорной вероятности.

Апостериорная вероятность может оцениваться по формуле Байеса. На практике это оказывается возможным, когда известны распределения вероятности показателей внутри каждого из классов. Такие распределения могут быть рассчитаны с помощью известного инструмента статистического оценивания — метода максимального правдоподобия. Однако для использования метода максимального правдоподобия необходимо сделать предположение о форме распределения. Обычно используется гипотеза о нормальном распределении. Распознавание с помощью байесовских классификаторов оказывается эффективным в той степени, в которой справедлива лежащая в их основе гипотеза о форме распределений.

Наивный байесовский классификатор [1.12]. По сравнению с общей схемой байесовской классификации вводится дополнительное предположение о статистической независимости показателей, по которым производится распознавание. Наивный байесовский классификатор (НБК) может применяться как для категориальных, так и для непрерывных переменных. При приме-

нении НБК к непрерывным переменным часто используется гипотеза о нормальном распределении. Многие авторы отмечают высокую эффективность НБК в задачах медицинской диагностики и прогнозирования [1.11; 1.28; 1.34]. При слабой взаимозависимости показателей НБК нередко превосходит альтернативные технологии.

Логистическая регрессия. В методе логистическая регрессия [1.12] аппроксимация апостериорной вероятности осуществляется с помощью логистической функции $\sigma(z) = \frac{1}{1+e^{-z}}$. График логистической функции представлен на рисунке 1.3. Из рисунка 1.3 видно, что логистическая функция устремляется к

Рис. 1.3. Логистическая функция

0 при отрицательных z меньше -4 и устремляется к 1 при положительных z выше 4. В основе метода логистической регрессии при распознавании целевого класса лежит идея использования логистической функции как приближения апостериорной вероятности. При этом величина z представляется в виде линейной комбинации показателей, по которым производится распознавание. Значения коэффициентов линейной комбинации подбираются таким образом, чтобы апостериорная вероятность, задаваемая логистической функцией, оказывалась как можно ближе к 1 для случаев из целевого класса и как можно ближе к 0 для случаев, не принадлежащих целевому классу. Обычно для этой цели используется метод максимального правдоподобия.

1.3.2. Метод опорных векторов

В основе метода опорных векторов [1.5; 1.12] лежит идея проведения линейной границы на равном удалении от двух распознаваемых классов, что для представленной на рисунке 1.4 задачи как раз соответствует сплошной ли-

нии. Рассмотрим сначала простую задачу распознавания, когда классы раз-

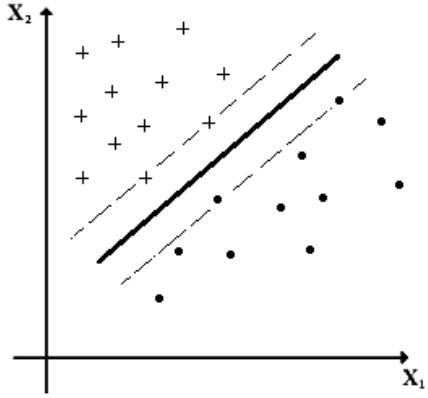

Рис. 1.4. Пример разделения двух классов границей, проходящей на равном удалении от каждого из них

деляются с помощью линейной границы. Очевидно, что данная задача решается неоднозначно. На рисунке приведён пример разделения объектов обучающей выборки из классов 1 и 2 по двум показателям. Линейными границами, разделяющими классы, на рисунке являются две пунктирные параллельные прямые линии и сплошная прямая, пролегающая посередине. Каждая из пунктирных линий максимально приближена к одному из классов так, что некоторые из объектов обучающей выборки лежат прямо на ней. Однако неоднозначность сохраняется из-за существования произвола в ориентации двух параллельных пунктирных границ.

Для того, чтобы сделать задачу совершенно однозначной выдвигается ещё и дополнительное требование ориентировать пунктирные прямые таким образом, чтобы зазор между ними был бы максимальным. В результате, проходящая посередине сплошная прямая линия окажется не только равноудалённой от классов, но и максимально удалённой от каждого из них. То есть, расстояние от этой прямой до ближайшего к ней объекта каждого из классов будет максимальным. Задача, представленная на рисунке, является двумерной. Для задач, когда классы разделяются по трём показателям, прямым граничным линиям соответствуют плоскости. Когда число показателей превышает 3, плоскости превращаются в гиперплоскости.

Поиск двух параллельных гиперплоскостей, разделяющих два класса, с максимальным зазором между ними сводится к хорошо изученной математи-

ками задаче квадратичного программирования. Существуют эффективные методы, позволяющие находить её решения для произвольных задач [1.5; 1.12].

Изложенная выше схема может быть использована только в тех случаях, когда классы линейно разделимы, то есть, когда разделяющая их линейная граница действительно существует. Однако авторами метода была предложена модификация метода, которая может быть использована также в случае, когда распознаваемые классы не являются линейно разделимыми. При этом две касающиеся разделяемых классов гиперплоскости подбираются таким образом, чтобы число ошибочных классификаций было минимальным. Наконец, в 1997 году в работе [1.5] была предложена ещё одна его модификация, которая позволила строить с его помощью не только линейные, но также и нелинейные разделяющие функции.

Метод опорных векторов используется для решения разнообразных медицинских задач. Например, в онкологии метод используется для диагностики типа опухолей по профилю экспрессии генов с помощью технологии микроэррэй (microarray) [1.21], оценки риска рецидива рака [1.13; 1.14]. В кардиологии метод опорных векторов использовался для диагностики типов аритмий по кардиограмме [1.26; 1.30].

1.3.3. Нейронные сети

В основе нейросетевых методов лежит попытка компьютерного моделирования процессов обучения, используемых в живых организмах. Когнитивные способности живых существ обеспечиваются функционированием сетей связанных между собой биологических нейронов – клеток нервной системы. Для моделирования биологических нейросетей используются сети расположенных в памяти компьютера искусственных нейронов - математических моделей отдельных нейронов. Обычно искусственные нейроны функционируют согласно схеме, представленной на рисунке 1.5. На вход нейрона по входным связям поступают сигналы, снимаемые с распознаваемых объектов или с выходов других нейронов. Каждой входной связи поставлен в соответствие ее вес. Искусственный нейрон включает в себя сумматор, вычисляющий сумму входных сигналов, умноженных на веса соответствующих

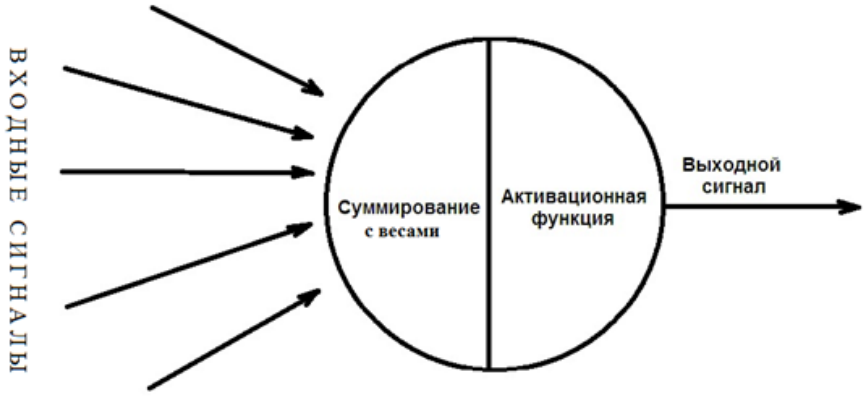

Рис. 1.5. Математическая модель нейрона

связей. Взвешенная сумма, вычисленная сумматором, далее преобразуется в выходной сигнал. В первой компьютерной модели нейрона, предложенной Ф. Розенблаттом, использовался простой бинарный преобразователь:

- выходной сигнал считался равным 1, если рассчитанная сумматором величина оказывалась положительной;
- выходной сигнал считался равным -1, если рассчитанная сумматором величина оказывалась отрицательной.

Настройка модели с целью достижения максимальной точности распознавания осуществлялась путём подбора оптимальных весов входных связей. Для этих целей использовался достаточно простой метод, в котором коррекция весов осуществлялась при ошибочной классификации вновь предъявляемого объекта. В современных нейроных сетях преобразование результатов действия сумматора в выходной сигнал осуществляется с помощью той же самой логистической функции, которая используется в методе логистическая регрессия.

Модель Ф. Розенблатта с некоторым успехом использовалась для решения задач распознавания в 60-70 годы прошлого века. Невозможность описания с её помощью нелинейных зависимостей привело к снижению интереса к методам нейронного моделирования по сравнению с давно известными

статистическими методами. Интерес возродился после создания компьютерных моделей, функционирующих как совокупность отдельных искусственных нейронов. Выяснилось, что такие модели позволяют описывать самые сложные нелинейные зависимости.

Утвердилась послойная архитектура построения нейронных сетей, представленная на рисунке 1.6.

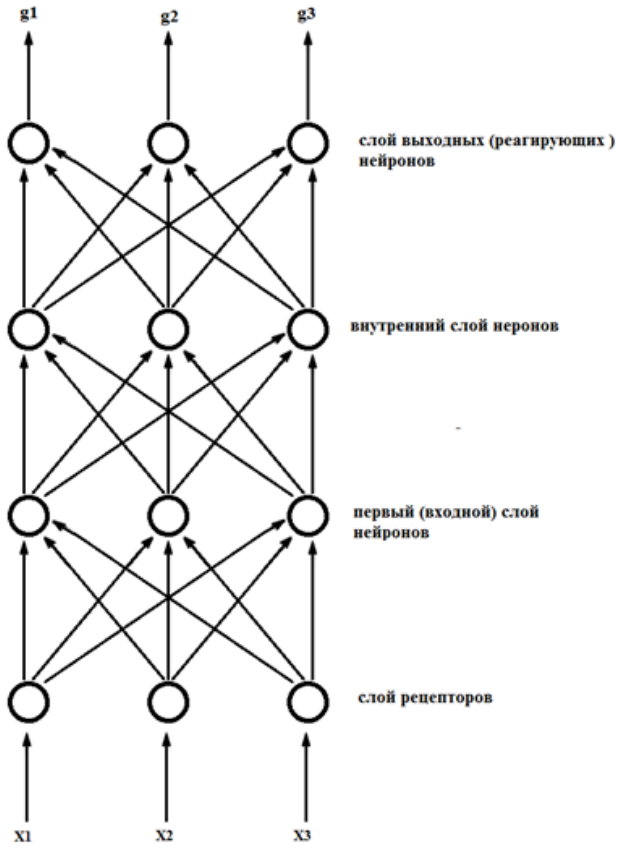

Рис. 1.6. Архитектура многослойной нейронной сети

На нейроны-рецепторы поступают сигналы (X_1, X_2, X_3), идущие от самих распознаваемых объектов. В качестве таких сигналов интерпретируются описывающие объекты показатели. Каждый рецептор преобразует соответствующий ему входной сигнал к виду, подходящему для дальнейшей обработки, и передаёт преобразованный сигнал на вход каждого из нейронов первого слоя. Выходные сигналы нейронов первого слоя подаются на вход нейронов

внутреннего слоя. Процесс заканчивается после вычисления выходных сигналов последнего слоя выходных нейронов (g_1, g_2, g_3), которые могут интерпретироваться как вероятности принадлежности распознаваемого объекта классам. Объект относится в наиболее вероятному классу. Слои нейронов, лежащие между первым и выходным слоями, принято называть внутренними слоями. На рисунке 1.5 такой слой только один, но их может быть несколько.

Обучение нейронной сети производится путём подбора оптимальных весов связей между искусственными нейронами из разных слоёв. Оптимизация весов производится с помощью многошаговой процедуры. На каждом шаге коррекция весов осуществляется по предъявленной порции вновь распознаваемых объектов. Направление, в котором производится коррекция соответствует наискорейшему снижению ошибки классификации для объектов из предъявленной порции. Обычно для коррекции весов используется метод обратного распространения ошибки (back propagation [1.12]), включающий эффективное определение направления наискорейшего уменьшения ошибки.

Всплеск интереса к нейронным сетям в последние годы тесно связан с развитием направления машинного обучения, называемого глубоким или глубинным обучением. Смысл глубинного обучения состоит в многостадийной обработке исходной информации, которая преобразуется к виду, удобному для принятия окончательного диагностического или прогностического решения. Например, на первом шаге ищется описание исходного изображения в виде набора значений показателей первого уровня. Показатели первого уровня преобразуются в более информативные показатели второго уровня, по которым и принимается окончательное решение. Очевидно, что такой многостадийный способ решения задач хорошо соответствует послойному преобразованию входной информации нейронной сетью.

Нейронные сети используются для решения разнообразных задач в различных областях медицины. При этом больший успех по сравнению с другими методами достигается при достаточно больших размерах обучающей информации. В качестве примера можно привести задачу диагностики рака груди по результатам маммографии с учётом включённых в модель демографических факторов и других характеристик. Нейросетевая модель строилась по

результатам обследования 466 злокачественных и 48267 доброкачественных новообразований. Использовалась нейронная сеть с большим количеством слоёв. Диагностическая способность разработанного метода оценивалась с помощью 10-блоковой кросс-валидации через величину AUC, которая составила 0,965 [1.2].

1.3.4. Решающие деревья и леса

Решающее дерево является алгоритмом распознавания, представляющим собой иерархически организованную систему вопросов, в результате ответов на которые произвольный распознаваемый случай может быть отнесён к одной из категорий (классов)[1.1; 1.12]. Решающее дерево может быть изображено графически в виде набора узлов, которые соединяются направленными стрелками. На рисунке графически представлен пример решающего дерева, предназначенного для диагностирования туберкулёза по рентгенограмме и другим показателям. Пример был взят из работы [1.1]. Решающее дерево было построено по выборке из 275 пациентов, для 27% из которых был установлен диагноз туберкулёз.Можно выделить корневой узел без входящей и с двумя выходящими стрелками. Выделяются также концевые узлы (листья), выделенные на рисунке более тёмным цветом, в которые входит одна стрелка, но выходящие стрелки отсутствуют. Все остальные узлы имеют одну входящую и две выходящие стрелки.

Вопросы соответствуют всем узлам решающего дерева, кроме листьев. В результате ответа на вопрос происходит переход к следующему узлу по направленной стрелке, соответствующей ответу. Каждому узлу решающего дерева соответствует совокупность диагностируемых случаев, которая формируется согласно ответам, полученным при прохождении предыдущих узлов. Корневой вершине соответствует исходная полная совокупность и наиболее информативный вопрос. Ответ на этот вопрос сам по себе позволяет получить наилучшее приближение решения рассматриваемой диагностической или прогностической задачи. В приведённом на рисунке примере корневой вершине соответствует вопрос: существуют ли на рентгенограмме изменения, за исключением остаточных?

В результате утвердительного или отрицательного ответа на этот вопрос выделяются две совокупности случаев, соответствующие второму и третьему узлу. Всего на обучающих данных второму узлу соответствует 85 случаев с отсутствием на рентгенограмме изменений или присутствием только остаточных изменений после перенесённой ранее болезни. Доля пациентов с туберкулёзом в этой группе составляет только 3.5%,что очевидно существенно ниже доли случаев с туберкулёзом во всей базе данных. Третьему узлу соответствует группа из 190 пациентов с наличием изменений, включая типичные, атипичные и возможные изменения. Доля пациентов с туберкулёзом в группе с изменениями составила 37.0%, что заметно выше доли пациентов с туберкулёзом во всей выборке. Второй узел является концевым.

Отнесение диагностируемого случая к данному узлу не предполагает задания дополнительных уточняющих вопросов. Второму узлу соответствует вопрос: являются ли изменения атипичными? Утвердительный ответ на этот вопрос приводит к пятому узлу, которому на обучающих данных соответствует 114 пациентов. Доля случаев с туберкулёзом среди них составляет 14%. Отрицательный ответ приводит к четвёртому узлу, которому на обучающих данных соответствует 76 пациентов. Доля случаев с туберкулёзом среди них составляет 71%. Процесс распознавания с помощью решающего дерева заканчивается по достижению концевого узла, которому соответствуют оценки вероятности классов. Например, при достижении концевого узла 2 вероятность туберкулёза составляет 3,5%, а при достижении концевого узла 7 вероятность туберкулёза составляет 78%.

Отметим, что для узлов, связанных с непрерывными показателями, задаётся вопрос о превышении показателем какого-то фиксированного порога. В приведённой на рисунке схеме непрерывным показателем является возраст пациента. Данный показатель соответствует узлу 8. Построение решающего дерева производится по обучающей выборке. На первом шаге формируется корневой узел. Для этого выбирается показатель и соответствующий вопрос, позволяющий выделить максимально однородные по содержанию распознаваемых классов выборки. Например, применение вопроса о наличии изменений на рентгенограмме позволяет выделить группу из 85 пациентов, в которой туберкулёз есть только у троих (3,5%). При анализе непрерывных показа-

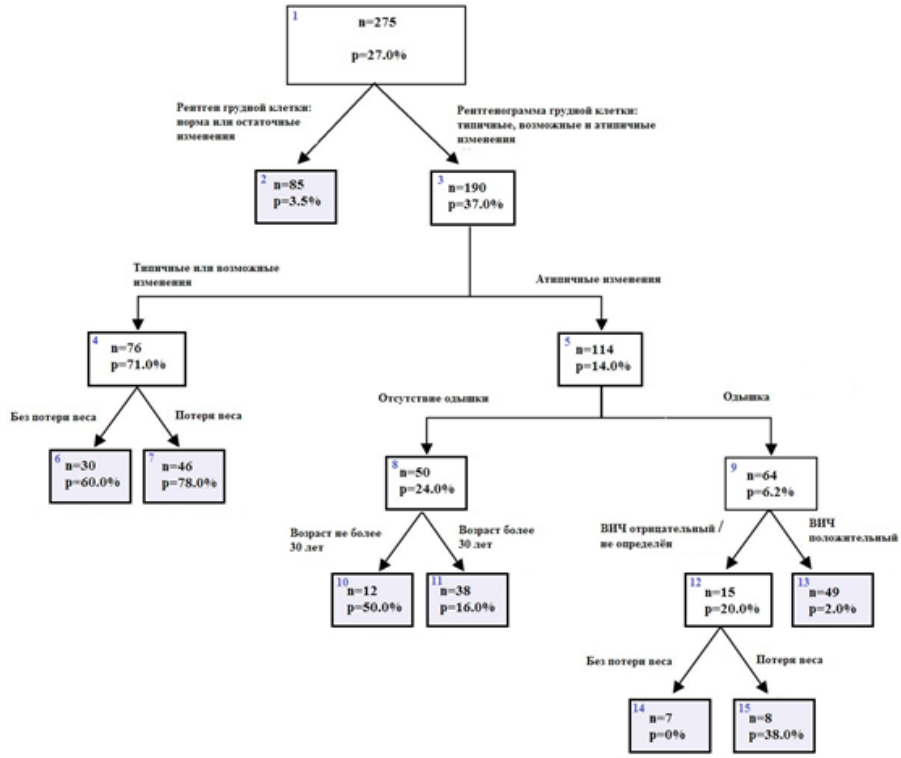

Рис. 1.7. Пример решающего дерева для диагностики туберкулёза (по материалам работы [1.1])

телей, производится поиск оптимального порогового значения соответствующего показателя. Например, в узле 8 для показателя возраст был выбран порог 30.

Далее ищутся показатели и вопрос, позволяющие оптимально разделить распознаваемые случаи в выборках, получившихся на первом шаге. Если для какой-либо из выборок не удаётся подобрать показатель и вопрос, позволяющий достоверно улучшить однородность во вновь образованных группах, то соответствующий этой выборке узел объявляется концевым. Так в решающем дереве, представленном на рисунке 1.7, концевым был объявлен левый узел второго уровня. По правому узлу ветвление было продолжено. Процесс построения дерева завершается, когда все достигнутые узлы оказываются концевыми.

1.4. Ансамбли алгоритмов

1.4.1. Эффективность коллективных решений

Мировой опыт использования методов машинного обучения показывает, что значительное повышение эффективности может быть достигнуто за счёт использования коллективов обученных алгоритмов . Такие коллективы в мировой научной периодике принято называть ансамблями. Коллективные прогнозы или диагностические решения вычисляются по прогнозам и диагностическим решениям, получаемым отдельными алгоритмами – членами ансамбля. Коллективные решения позволяют не только точнее аппроксимировать зависимости, объективно существующие в данных, но и достигать существенно более высокой устойчивости по отношению к чисто случайным изменениям. Для достижения высокой эффективности ансамбля алгоритмов необходимо, чтобы входящие в него алгоритмы максимально отличались друг от друга. В этом случае правильные решения одного алгоритма будут компенсировать ошибочные результаты другого алгоритма. Участие в ансамбле большого числа разнообразных алгоритмов приводит к увеличению вероятности получить большинство правильных ответов.

Для формирования оптимального ансамбля могут быть использованы алгоритмы, построенные с помощью различных методов машинного обучения, в том числе упомянутых выше. Однако такой подход является достаточно трудоемким и не позволяет получить ансамбль с большим количеством алгоритмов-«участников». Альтернативным способом является использование алгоритмов, работающих на различных выборках. Такие выборки обычно формируются из исходной обучающей выборки с помощью так называемых процедур ресемплинга [1.19] . В результате процедуры ресемплинга получается измененная выборка с тем же количеством объектов, в которой одни объекты могут повторяться, а другие вообще отсутствуют по сравнению с исходной выборкой. В случае, когда выборки генерируются независимо, то процедура формирования ансамбля алгоритмов носит название «бэггинг» (bagging) [1.12; 1.19].

Наряду с этими процедурами используют процедуру генерации новых выборок, в которых при генерации используется информация о результатах распознавания на предыдущих сгенерированных выборках. При этом в каждой последующей выборке увеличивается число объектов, которые были ошибочно классифицированы ранее. Такой метод генерации используется в процедурах, которые носят название «бустинг» (boosting). Высокая эффективность распознавания достигается, если методы бэггинг и бустинг применять в сочетании с методом решающих деревьев. В этом случае формируются ансамбли решающих деревьев, которые называют решающими лесами (random forests) [1.12; 1.19].

1.4.2. Решающие леса

Метод случайных решающих лесов был с успехом использован для ранней диагностики болезни Альцгеймера по томографическим изображениям, полученным с помощью Однофотонной эмиссионной компьютерной томографии (далее — ОЭКТ)[1.22]. Обучающая выборка включала ОЭКТ изображения головного мозга 56 пациентов, имеющих симптомы болезни Альцгеймера различной степени выраженности, и 41 пациента без симптомов. Распознавание двух групп пациентов проводилось по показателям, извлечённым из изображений с помощью различных технологий. Применение метода случайных решающих лесов позволило достичь чувствительности 100% при специфичности 92.6%. Оценка точности проводилась методом скользящего контроля.

1.4.3. Ансамбли закономерностей

Среди коллективных методов можно выделить группу методов, основанных на принятии коллективных решений по системам эмпирических закономерностей различного типа . Так, под закономерностью может пониматься связь целевой переменной с условиями, накладываемыми на показатели, как это будет описано далее.

На рисунке 1.8, который взят из работы [1.17], приведён пример закономерности, связывающей частоту возникновения транзиторной ишемической атаки (далее — ТИА) у пациентов пожилого возраста с хронической

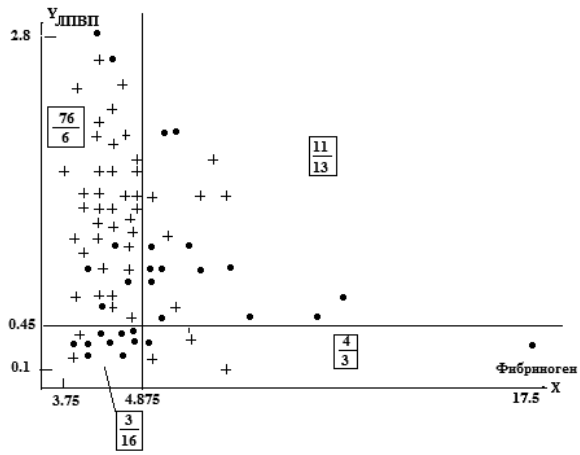

Рис. 1.8. Двумерная диаграмма рассеяния, полученная с помощью метода оптимально достоверных разбиений из работы [1.17]

ишемией головного мозга с уровнем фибриногена и ХС-ЛПВП. Значком «+» помечены случаи без ТИА, значком «о» случаи с ТИА. Превышение ХС-ЛПВП уровня 0,45 ммоль на диаграмме соответствует верхняя правая прямоугольная область, выделяемая границами, параллельными координатным осям. Статистика соответствующих этой области случаев с ТИА и без ТИА приведена в виде дроби, в числителе которой даётся число случаев, помеченных «+», а в знаменателе приводится число случаев, помеченных «о».

Таким образом, из 82 случаев, соответствующих верхней левой области, к классу с ТИА относится только 6 значений. Поэтому об этой области можно говорить как о закономерности, связанной с отсутствием ТИА. Напротив, в нижней левой области из 19 соответствующих ей случаев для 16 отмечена ТИА. Приведённая на рисунке закономерность является двумерной. Однако могут быть использованы закономерности с более высокой размерностью. Закономерности, задаваемые границами, параллельными координатным осям, по сути дела задаются как конъюнкции условий, наложенных на отдельные показатели. Поэтому такие закономерности принято называть логическими. Наряду с логическими закономерностями используются также двумерные закономерности с границами, произвольно ориентированными относительно координатных осей.

Как правило, по данным можно построить достаточно большое число закономерностей. При поступлении нового объекта, для которого необходи-

мо провести диагностику или сделать прогноз, проверяется, в какие закономерности попали значении описывающих его показателей. Для объекта может прогнозироваться значение целевой переменной, которому соответствует большинство таких закономерностей. То есть, может использоваться простое голосование по большинству. Может использоваться также способ вычисления коллективных решений, учитывающий значимость закономерностей.

Например, при вычислении коллективного решения большим весом обладают закономерности, включающие большое число объектов из обучающей выборки. Важна также выраженность закономерности, то есть степень преобладания внутри закономерности соответствующего ей значения целевой переменной. Существует целый ряд подходов к поиску закономерностей и вычислению коллективных решений [1.32; 1.33]. В методе «Логические закономерности» используются многомерные логические закономерности, содержащие внутри себя объекты только одного класса и допускающие минимальное попадание объектов других классов.

1.4.4. Метод мультимодельных статистически взвешенных синдромов

В методе Мультимодельные статистически взвешенные синдромы (далее — МСВС) [1.16; 1.27] используются только закономерности с размерностью не выше двух. Однако, используются также двумерные закономерности с границами, произвольно ориентированными относительно координатных осей.

Методы, основанные на голосовании по системам закономерностей, нередко оказываются весьма успешными, превосходя альтернативные подходы. В качестве примера можно привести задачу прогнозирования: оценить риск возникновения новых осложнений в первые полгода после перенесенного обострения ишемической болезни сердца [1.9]. Исследования проводились по выборке, содержащей значения 407 разнообразных показателей для 1193 пациентов. У 136 из них в течение полугода наблюдались осложнения. С помощью метода отбора признаков, основанного на фильтрации, был отобран 41 информативный показатель. Метод МСВС позволил достигнуть

величины AUC= 0,72, что оказалось заметно выше результатов, достигнутых с помощью альтернативных технологий [1.9].

1.5. Методы отбора признаков

Значительно повысить эффективность решения задач диагностики и прогнозирования позволяют методы выделения в исходном наборе совокупности показателей, при использовании которых достигается максимальная точность. Остальные показатели при этом игнорируются. Возможность повышения точности при сужении набора показателей связана с существенным повышением устойчивости настройки моделей благодаря удалению шумовых или малозначимых показателей. Важность отбора переменных повышается в задачах, где исходная размерность высока. В качестве примера можно привести задачи, основанные на использовании профиля экспрессии генов с помощью технологии микроэррэй (microarray) [1.3].

В настоящее время можно выделить три основных направления отбора информативных показателей: методы, основанные на фильтрации: «обёрточные методы» (wrapper methods) и методы со встроенным отбором [1.10]. В методах, основанных на фильтрации, показатели оцениваются по отдельности. То есть отбираются признаки, значения которых в распознаваемых классах существенно различаются. Для отбора таких признаков могут использоваться, например, разнообразные статистические критерии (критерий Хи-квадрат, Стьюдента, Манна-Уиттни и др.). Также может оцениваться возможность разделения распознаваемых классов с помощью порогового правила для одного показателя. В качестве меры информативности показателя при этом используется величина AUC. Преимуществом методов, основанных на фильтрации, является возможность их эффективного использования при очень большом числе исходных показателей. Недостатком фильтрации является невозможность учёта при их использовании эффектов, связанных с взаимодействием признаков.

«Обёрточные» методы состоят в подборе таких сочетаний признаков, для которых достигается наибольшая точность диагностики или прогнозирования. Например, может осуществляться поиск такой комбинации, при

использовании которой точность диагностики при оценивании её методом скользящего контроля является максимальной. Недостатком таких методов является неустойчивость отбора при большом числе исходных показателей и ограниченном объёме обучающих данных.

Наконец, многие модели машинного обучения включают отбор признаков, автоматически исключая неинформативные признаки из генерируемых алгоритмов. Такой способ отбора принято называть встроенным. Из упомянутых в настоящей статье к методам со встроенным отбором могут быть отнесены методы решающих деревьев и лесов, а также методы, основанные на ансамблях закономерностей. Одним из известных методов отбора переменных является эластичная сеть. Данный метод основан на использовании так называемых штрафов за слишком большие значения параметров, характеризующих вклад в алгоритм распознавания отдельных признаков. По таким параметрам производится настройка (обучение) алгоритмов распознавания. Математически было показано, что при использовании таких штрафных функций значения параметров, соответствующих малозначимым признакам, оказывается равным 0, что на самом деле означает отказ от их использования.

В работе [1.20] метод эластичная сеть был использован для отбора информативных показателей при оценивании риска летального исхода для пациентов, помещённых в бокс интенсивной терапии, по тексту, который записывался наблюдающим специалистом. Известные методов анализа текстов позволили вычислить 1842522 исходных показателей. Использование метода эластичной сети позволило отобрать только 465 из них. В результате отбора признаков эффективность по AUC увеличилась с 0.791 до 0.889.

1.6. Байесовские сети

В отличие от рассмотренных ранее методов машинного обучения, предназначенных для прогнозирования одной целевой переменной по набору признаков, байесовские сети позволяют описать взаимозависимость всех показателей, относящихся к некоторой рассматриваемой задаче. Байесовская сеть включает в себя граф, вершинам которого соответствуют отдельные переменные. Вершины графа связаны между собой направленными рёбра-

ми (стрелками). Такой граф называется ориентированным. Если стрелка направлена из вершины Р в вершину А, то вершина Р называется предком вершины А, а вершина А называется потомком вершины Р. Граф байесовской сети характеризует взаимную зависимость переменных, соответствующих его вершинам.

Наряду с ориентированным графом байесовская сеть для каждой переменной Р включает также информацию о вероятности различных значений Р в зависимости от переменных, являющихся предками Р. Такая информация может задаваться в виде математических функций или в табличном виде. Преимуществом байесовской сети является возможность описания сложной вероятностной зависимости в максимально компактном виде. Пример достаточно простой байесовской сети, описывающей взаимосвязь нескольких важных прогностических факторов при внезапной остановке сердца, представлен в работе [1.15]. Целью моделирования было описание связи вероятности летального исхода с полом; возрастом; этиологией; типом сопутствующего нарушения сердечного ритма (НСР), выявляемого на кардиограмме, измеренной непосредственно после остановки сердца (НСР-ЭКГ); действиями спасателя (ДС), оказавшегося в этот момент рядом с больным.

В рамках разработанной модели принималось, что возраст имеет три уровня градации: до 45 лет, от 45 до 60 лет и свыше 60 лет. Выделялись две возможные причины внезапной остановки: кардиальная причина, связанная с заболеванием самого сердца; некардиальная причина, связанная с иными заболеваниями. Рассматривались следующие типы нарушения сердечного ритма:

- фибрилляция желудочков — дезорганизованная электрическая активность миокарда желудочков без механической активности;
- желудочковая тахикардия с отсутствием пульса на крупных сосудах;
- электрическая активность без пульса;
- асистолия, то есть отсутствие электрической активности желудочков.

К действиям спасателя относились комбинирование искусственного дыхания с непрямым массажем сердца, а также использование дефибриллятора. Восстановление спонтанной циркуляции крови является непосредственным фактором выживания больного.

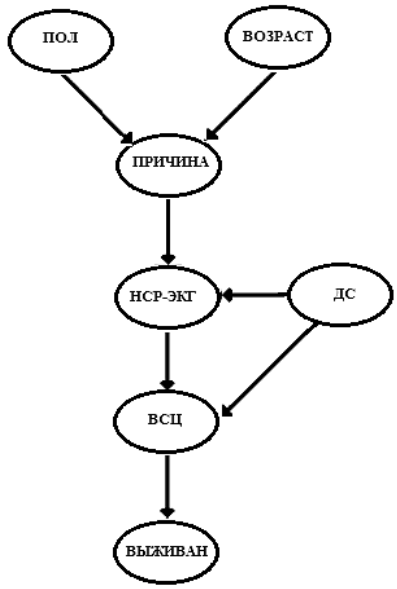

Рис. 1.9. Граф байесовской сети для задачи оценивания вероятности летального исхода при внезапной остановке сердца (по материалам работы[1.15])

Из графа байесовской сети (рисунок 1.9) видно, что факторами, определяющими вероятность каждой из причин внезапной остановки сердца, являются пол и возраст. На вероятность типа нарушения сердечного ритма оказывают непосредственное влияние причина внезапной остановки и действия спасателя. Вероятность восстановления спонтанной циркуляции непосредственно связаны с типом нарушения и действиями спасателя. Очевидно, что восстановление спонтанной циркуляции зависит от пола и возраста, но согласно модели эта связь происходит только через тип НСР. Иными словами, при известном типе НСР вероятность восстановления от пола и возраста не зависит.

Структура графа байесовской сети может задаваться экспертом, что существенно облегчает задачу вычисления вероятностной связи каждого показателя с показателями, являющимися его предками. За последние годы были разработаны эффективные технологии построения ориентированного графа и всей байесовской сети непосредственно из данных. Байесовские сети используются для решения задач диагностики в различных областях медицины. Например, можно привести задачу диагностики рака груди по цитологическим показателям [1.6]. Особый интерес вызывают области применения,

связанные с использованием генетической информации. В работе исследовалась связь риска рака мочевого пузыря с однонуклеотидным генным полиморфизмом, факторами окружающей среды и другими показателями.

1.7. Информативная визуализация

Алгоритмы диагностики или прогнозирования, генерируются с помощью методов МО в автоматическом режиме. При практическом использовании такие алгоритмы нередко функционируют в режиме «чёрного ящика». Пользователь подаёт на вход алгоритма значения используемых показателей, а на выходе получает диагностическое или прогностическое решение без понимания причин, по которым оно было предложено. При использовании технологий МО в медицине режим «чёрного ящика» оказывается существенным недостатком, так как непонимание причин негативного прогноза не позволяет эффективно выбрать стратегию, направленную на его улучшение.

Следует отметить, что высокая прозрачность принимаемых решений обеспечивается байесовскими сетями и решающими деревьями. Однако эффективное обучение байесовских сетей достигается только при больших размерах обучающей информации. Альтернативным способом преодоления проблемы «чёрного ящика» является метод информативной визуализации, заключающийся в локализации диагностируемого случая на диаграммах рассеяния, соответствующих парам информативных показателей, используемых сгенерированным диагностическим или прогнозным алгоритмом.

На предварительном этапе с использованием метода оптимальных достоверных разбиений (ОДР) для пар информативных показателей выделяются области соответствующие различным вариантам прогноза [1.17]. Причем ранжирование по информативности позволяет обратить внимание аналитика на самые статистически значимые и информативные с точки зрения прогноза показатели. Пользователь может в интерактивном режиме просмотреть положение прогнозируемого случая на диаграммах и оценить, какие именно сочетания вызывают смещение прогноза в ту или иную сторону.

На рисунке 1.10 представлен пример подхода информативной визуализации с помощью приложения Data Master Azforus к известной задаче Hepatitis

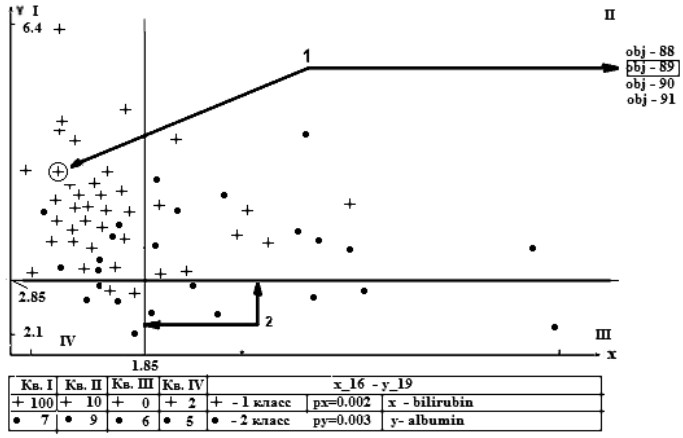

Рис. 1.10. Представлено использование информативной визуализации применительно к задаче Hepatitis Domain

Domain, взятой из UCI репозитория [1.29]. Целью является прогнозирование исхода заболевания по набору из 18 показателей. Представлена диаграмма рассеяния для содержания в крови билирубина (ось *X*) и альбумина (ось *Y*). Методом оптимально достоверных разбиений поставлены границы таким образом, что в каждом базовом множестве (квадранте) преобладают значения какого-то одного класса (стрелки 3). Кружками обозначены случаи с летальным исходом вследствие гепатита, а крестики соответствуют выживанию пациента.

Из рисунка видно, что попадание значений билирубина и альбумина в верхний левый квадрант (I) свидетельствует о большей вероятности выживания, поскольку данному квадранту соответствует 100 случаев с выживанием и только 7 летальных исходов. Наоборот, попадание значений содержания билирубина и альбумина в нижний правый квадрант (III) свидетельствует о высокой вероятности летального исхода, поскольку данному квадранту соответствует 6 летальных исходов и не соответствует случаев с выживанием.

Данный подход помогает акцентировать внимание именно на самых информативных показателях с точки зрения разделения двух классов при диагностике или прогнозировании. Стрелка 2 указывает на список информативных пар показателей, ранжированных по функционалу Хи-квадрат. Значения функционала *F* находятся в соседнем столбце. Аналогичное ранжиро-

вание возможно по весам за класс, что тоже, в свою очередь, позволит аналитику среди большого числа сочетаний пар показателей выделить самые важные. Над этими столбцами расположены названия объектов (ФИО пациентов). При активировании объекта в списке на диаграмме кружком будет обозначено соответствующее значение (стрелка 1). Для врача может быть важно соседство пациентов с общей симптоматикой заболевания в многомерном пространстве признаков. Можно создавать цепочку «соседей», ведущую по базовым множествам из класса с плохим прогнозом в класс с хорошим прогнозом. При этом первыми будут показаны именно показатели с наиболее высокими весами, т.е. высоко значимые с точки зрения разделения классов.

Выводы

Из приведённого обзора можно сделать вывод о существенном прогрессе в развитии методов машинного обучении, достигнутом за последние годы. Возрастает интенсивность медико-биологических исследований, связанных с использованием МО в медицине. Такой рост связан с накоплением большого объема самых разнообразных по своей природе данных, с потребностью их анализа и возможностью привлечения большого числа методов для получения диагностических или прогнозных решений.

Выявленные в результате машинного обучения закономерности представляют собой материал для дальнейших исследований в целях детального понимания соответствующих биологических (или любых других) механизмов. Рост популярности методов МО зависит также от повсеместного распространения электронных историй болезни, что позволяет формировать базы данных большого размера, а также активное развитие конкурирующих друг с другом технологий МО, что позволяет выбирать метод наиболее подходящий для решения каждой конкретной задачи. Методы МО позволяют принимать врачебные решения с учётом большого числа различных факторов, что делает их ориентированными на каждого конкретного пациента.

Благодарности

Статья была подготовлена при поддержке РФФИ (грант №20-01-00609).

Библиографический список

1.1. *Aguiar F.*, *et.al.* Classification and regression tree (CART) model to predict pulmonary tuberculosis in hospitalized patients // BMC Pulmonary Medicine. — 2012. — Vol. 12. — P. 1–40.

1.2. *Ayer T.*, *et.al.* Breast cancer risk estimation with artificial neural networks revisited // Cancer. — 2010. — Vol. 4. — P. 465–504.

1.3. *Boulesteix A. L.*, *Strobl C. T.*, *Augustin M.* Evaluating microarray-based classifiers: an overview // Cancer Inform. — 2008. — Vol. 6. — P. 77–97.

1.4. *Chrominski K.*, *Tkacz M.* Comparison of Outlier Detection Methods in Biomedical Data // Journal of Medical Informatics & Technologies. — 2010. — Vol. 16. — P. 89–94.

1.5. *Cortes C.*, *Vapnik V.* Support-vector networks. Vol. 20. — Machine Learning, 1995. — 273 p.

1.6. *Cruz-Ramirez N.*, *Acosta-Mesa H. G.*, *et.al.* Diagnosis of breast cancer using Bayesian networks: a case study // Comput Biol Med. — 2007. — Vol. 37. — P. 1553–64.

1.7. *Darcy A.*, *Louie A.*, *Roberts L.* Machine Learning and the Profession of Medicine // Journal of American Medical Association. — 2016. — Vol. 315, no. 6. — P. 551–552.

1.8. *Fawcett T.* An introduction to ROC analysis // Pattern Recognition Letters. — 2006. — Vol. 27. — P. 861–874.

1.9. *Guliev R.*, *et.al.* The use of partitionings for multiparameter data analysis in clinical trials // Matematicheskaya Biologia i Bioinformatika. — 2016. — Vol. 11. — P. 46–63. — (In Russian).

1.10. *Guyon I.*, *Elisseeff A.* An Introduction to Variable and Feature Selection // Journal of Machine Learning Research. — 2008. — Vol. 3. — P. 1157–1182.

1.11. *Han J.*, *Kamber M.*, *Pei J.* Data mining: concepts and techniques. — 3rd ed. — Elsevier, 2012. — 774 p.

1.12. *Hastie T.*, *Tibshirani R.*, *Friedman J.* Statistical Learning. Data Mining, Inference, and Prediction. — 2nd ed. — Springer, 2008. — 550 p.

1.13. *Kim K.*, *Lee J.*, *et.al.* Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine // J Breast Cancer. — 2012. — Vol. 18. — P. 230–238.

1.14. *Kourou K. Exarchos T. P.*, *et.al.* Machine learning applications in cancer prognosis and prediction // Computational and Structural Biotechnology Journal. — 2015. — Vol. 13. — P. 8–17.

1.15. *Krizmaric M.*, *Mertik M.* Application of Bayesian networks in emergency medicine // Journal of Machine Learning Research. — 2009. — Vol. 7. — P. 1157–1182.

1.16. *Kuznetsov V. A.*, *Kuznetsova A. V.*, *et.al.* Syndrome Approach for Computer Recognition of Fuzzy Systems and its Application to Immunological Diagnostics and Prognosis of Human Cancer // Journal of Chemical Physics. — 1996. — Vol. 15. — P. 81–100.

1.17. *Kuznetsova A. V.*, *Kostomarova I. V.*, *Senko O. V.* Logical and Statistical Analysis of Relationship Between Clinical and Laboratory Indices and Disturbance of Blood Circulation in Elderly Patients with Chronic Ischemia of the Brain // Matematicheskaya Biologia i Bioinformatika. — 2013. — Vol. 8. — P. 182–224. — (In Russian).

1.18. *Kuznetsova A. V.*, *Senko O. V.* Possibilities to use Data Mining techniques in medical and laboratory studies for discovering regularities in data // Information technologies and the Physician. — 2005. — No. 2. — P. 49–56. — (In Russian).

1.19. *Malley J. D.*, *Malley K. G.*, *Pajevic S.* Statistical Learning for Biomedical Data. — Cambridge University Press, 2011. — 281 p.

1.20. *Marafino B. J.*, *Boscardin W. J.*, *Dudley R. A.* Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes // J Biomed Inform. — 2015. — Vol. 54. — P. 114–20.

1.21. *Ramaswamy S.* Multiclass cancer diagnosis using tumor gene expression signatures // Proc. Natl Acad. Sci. USA 98. — 2001. — P. 15149–15154.

1.22. *Ramirez J.*, *et.al.* Computer aided diagnosis system for the Alzheimer's disease based on partial least squares and random forest SPECT image classification // Neuroscience Letters. — 2010. — Vol. 472. — P. 99–103.

1.23. *Rosenblatt F.* The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain // Cornell Aeronautical Laboratory, Psychological Review. — 1958. — Vol. 65, no. 6.

1.24. *Saar-Tsechansky M.*, *Provost F.* Handling Missing Values when Applying Classification Models // Journal of Machine Learning Research. — 2007. — Vol. 8. — P. 1625–1657.

1.25. *Samuel A.* Some Studies in Machine Learning Using the Game of Checkers // IBM Journal. — 1959. — Vol. 3, no. 3. — P. 201–213. — (In Russian).

1.26. *Sansone M.*, *et.al.* Electrocardiogram Pattern Recognition and Analysis Based on Artificial Neural Networks and Support Vector Machines // A Review Journal of Healthcare Engineering. — 2013. — Vol. 4. — P. 465–504.

1.27. *Senko O. V.*, *Kuznetsova A. V.* A recognition method based on collective decision making using systems of regularities of various types // Pattern Recognition and Image Analysis. — 2010. — Vol. 20. — P. 152–162.

1.28. *Subbalakshmi G.*, *Ramesh K.*, *Rao M.* Decision support in heart disease prediction system using naive Bayes // Indian J Comput Sci Eng. 20 Healthc Inform Res. — 2016. — Vol. 22. — P. 196–205.

1.29. The UC Irvine Machine Learning Repository. Hepatitis. — URL: https://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis (visited on 06.03.2019).

1.30. *Wasan P.*, *et.al.* Application of statistics and machine learning for risk stratificationof heritable cardiac arrhythmias // Expert Systems with Applications. — 2013. — Vol. 40. — P. 2476–2486.

1.31. *Zhuravlev Y. I.* About algebraic approach in recognition and classification tasks // Cybernetics Problems. Vol. 33. — 1978. — P. 5–68. — (In Russian).

1.32. *Zhuravlev Y. I.*, *Nazarenko G. I.*, *et.al.* Methods for discrete analysis of medical information based on recognition theory and some of their applications // Pattern Recognition and Image Analysis. — 2012. — Vol. 3. — P. 17–20.

1.33. *Zhuravlev Y. I.*, *Ryazanov V. V.*, *Senko O. V.* Recognition. Mathematical Methods. Program System. Applications. — Phazis, 2006. — 159 p. — (In Russian).

1.34. *Zmiri D.*, *Shahar Y.*, *Taieb-Maimon M.* Classification of patients by severity grades during triage in the emergency department using data mining methods // J Eval Clin Pract. — 2012. — Vol. 18. — P. 378–88.

1.35. *Zweig M. H.*, *Campbell G.* Receiver-Operating Characteristic(ROC) Plots: a Fundamental Evaluation Tool in Clinical Medicine // Clin. Chem. — 1993. — Vol. 39, no. 4. — P. 561–577.