Министерство образования Республики Беларусь

Учреждение образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Кафедра интеллектуальных информационных технологий

М. Д. СТЕПАНОВА, С. А. САМОДУМКИН, Т. Л. ЛЕМЕШЕВА

МАТЕМАТИЧЕСКИЕ МЕТОДЫ ДИАГНОСТИКИ В МЕДИЦИНСКИХ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМАХ

УЧЕБНО – МЕТОДИЧЕСКОЕ ПОСОБИЕ

по курсу "Прикладные интеллектуальные системы и системы принятия решений" для студентов специальности Т 10.04.00 "Искусственный интеллект"

Минск 2001

УДК 007:159.955: 519.23/.24 (075.8)

ББК 32.813 я 73

С 79

Степанова М. Д. , Самодумкин С. А., Лемешева Т. Л.

С 79 Математические методы диагностики в медицинских интеллектуальных системах: Учебно-методическое пособие по курсу "Прикладные интеллектуальные системы и системы принятия решений" для студентов специальности Т 10.04.00 "Искусственный интеллект". – Мн.: БГУИР, 2001. – 44 с.

ISBN 985 - 444 – 226 - 8

В методическом пособии формулируются проблемы классификации (распознавания), дается постановка задачи диагностики как классификационной задачи. Рассматриваются детерминированные, вероятностно–статистические, логические и структурные методы и соответствующие им математические модели, используемые в интел-лектуальных медицинских системах для решения диагностических задач. Приводятся примеры, иллюстрирующие применение методов классификации.

Пособие предназначено для курса ‴"Прикладные интеллектуальные системы и системы принятия решений".

ISBN 985 - 444 – 226 - 8

СОДЕРЖАНИЕ

1. ЗАДАЧИ КЛАССИФИКАЦИИ

Основная цель классификации (распознавания) состоит в построении правила, в соответствии с которым устанавливается, к какому из классов объектов (образов) может быть отнесен классифицируемый (распознаваемый) объект. При этом под классом понимается некоторая совокупность (подмножество) объектов, обладающих близкими свойствами. Методы классификации (распознавания образов) нашли широкое применение в прикладных системах различного назначения: медицина, техническая диагностика, геология и разведка полезных ископаемых, аэрокосмические методы исследований, робототехника, криминалистика и т. д. Использование методов классификации в медицинской диагностике и создание на их основе компьютерных систем позволяет изменить эффективность диагностики за счет увеличения числа анализируемых симптомов, оперативности, точности и достоверности.

Каждая прикладная система классификации предназначена для классификации объектов и явлений конкретной предметной области и требует построения математической модели. Рассмотрим основные задачи, возникающие в процессе проектирования систем классификации (распознавания).

1.1. Основные задачи классификации (распознавания)

Задача 1 заключается в определении полного перечня признаков (параметров), характеризующих объекты или явления, для классификации которых разрабатывается данная система. Признаки могут быть подразделены на детерминированные, вероятностные, логические и структурные.

Детерминированные признаки – это признаки, принимающие конкретные числовые значения. При рассмотрении детерминированных признаков ошибками измерений пренебрегают.

Вероятностные признаки – это признаки, случайные значения которых распределены по всем классам объектов, при этом решение о принадлежности распознаваемого объекта к тому или иному классу может приниматься только на основании конкретных значений признаков данного объекта, полученных в результате проведения соответствующих экспериментов. Признаки следует рассматривать как вероятностные и в том случае, когда измерение их числовых

значений производится с такими ошибками, что по результатам измерений невозможно определенно сказать, какое числовое значение данная величина приняла.

Логические признаки объектов можно рассматривать как элементарные высказывания, принимающие два значения истинности ("да", "нет" или "истина", "ложь") с полной определенностью. К логическим признакам относятся прежде всего признаки, не имеющие количественного выражения. Эти признаки представляют собой суждения качественного характера типа наличия или отсутствия некоторых свойств или некоторых элементов у распознаваемых объектов или явлений. В качестве логических признаков можно рассматривать, например, наличие боли, наличие в прошлом определенных заболеваний и т. д. К логическим признакам можно отнести такие признаки, у которых важна не сама величина, а лишь факт попадания или непопадания ее в заданный интервал.

Структурные (лингвистические, синтаксические) признаки представляют собой непроизводные элементы (символы) структуры объекта. Иначе эти элементы (константы) называют терминалами. Каждый объект может рассматриваться как цепочка терминалов или как предложение. При этом если предложение, описывающее неизвестный распознаваемый объект, относится к языку данного класса, то выносится решение о принадлежности объекта этому классу.

Все перечисленные виды признаков, описывающих объекты классификации, могут появляться при задании исходных данных в одной из следующих форм либо их сочетаниях: 1) экспертные данные, численная и символьная информация общего вида; 2) полученные в различных частях спектра излучений изображения (оптические, инфракрасные, ультразвуковые и т. д.) и затем преобразованные в цифровую форму; 3) сигналы (длинные числовые последовательности).

Задача 2 состоит в формировании априорной совокупности признаков. С учетом результатов решения **задачи 1** в эту совокупность включаются только те признаки, относительно которых может быть получена априорная информация, необходимая для описания классов с помощью этих признаков.

Задача 3 состоит в описании всех классов на языке признаков.

Если признаки классифицируемых объектов – детерминированные, то описанием каждого класса объектов на языке этих признаков является его

эталон, т. е. точка, сумма расстояний которой от точек, описывающих объекты, принадлежащие данному классу, минимальна.

Если признаки классифицируемых объектов – логические и имеют количественные выражения, то для описания классов объектов на языке признаков необходимо определить диапазоны значений признаков, соответствующие классам Ω_i, i = 1, …, m. При этом каждый из отрезков может рассматриваться как элементарное логическое высказывание A, B, C,…. Если признаки распознаваемых объектов суть суждения качественного характера, то каждый из них также рассматривается как элементарное логическое высказывание $A', B', C', \ldots$. Для описания классов на языке этих признаков необходимо выяснить, какими из них характеризуется каждый класс, после этого установить зависимости в форме булевых соотношений между признаками A, B, C,…, A', B', C' и классами Ω_i, i = 1, …, m.

Если распределение объектов по областям D_i p – мерного пространства признаков для всех значений i = 1, …, m вероятностное, то для описания классов необходимо определить характеристики этих распределений: функции плотности вероятности $f_i(x_1, x_2, \ldots x_p)$ значений признаков $x_1, x_2, \ldots x_p$ при следующем условии: априорные вероятности того, что случайным образом выбранный из общей совокупности объект окажется принадлежащим Ω_i, равны $P(\Omega_i)$.

Если признаки классифицируемых объектов – структурные, то описаниями классов являются языки, состоящие из предложений, каждое из которых характеризует структурные особенности объектов, принадлежащих исключительно одному из классов.

Задача 4 заключается в разбиении априорного пространства признаков на области, соответствующие классам. Подобное разбиение должно быть выполнено в некотором смысле оптимальным образом, например, так, чтобы при этом обеспечивалось минимальное значение ошибок, возникающих при классификации неизвестных объектов или явлений.

Пусть объект (наблюдение) описывается вектором $x = (x_1, x_2, \ldots x_p)$, состоящим из p компонентов – признаков. Тогда в пространстве признаков каждый объект будет представлять p – мерную точку.

Положим, произведено разбиение объектов на классы Ω_i (i = 1, …, m). Требуется выделить в пространстве признаков области D_i, эквивалентные классам: если объект, имеющий признаки x_j^0, j =1, …, p, относится к классу Ω_i, то представляющая его точка в пространстве признаков принадлежит области D_i.

Помимо геометрической существует и алгебраическая трактовка задачи: требуется построить разделяющие функции $F_i(x_1, x_2, ...x_p)$, i = 1, …, m, обладающие следующим свойством: если объект, имеющий признаки x_j^0, j = 1, …, p, относится к классу Ω_i, то величина $F_i(x_1^0, x_2^0, ...x_p^0)$ должна быть наибольшей. Если x_k обозначает вектор признаков объектов, относящихся к классу Ω_k, то $F_k(x_k) > F_l(x_k)$, k, l = 1, …, m, $k \neq l$.

Таким образом, в пространстве признаков **граница разбиений, называемая решающей границей между областями D_i, соответствующими классам Ω_i, выражается уравнением $F_k(x) - F_l(x) = 0$.**

Задача 5 состоит в выборе алгоритмов классификации (распознавания), обеспечивающих отнесение классифицируемого объекта или явления к тому или другому классу.

Алгоритмы классификации основаны на сравнении меры близости или меры сходства классифицируемого объекта с каждым классом. При этом если выбранная мера близости данного объекта с каким – либо классом Ω_k, k =1, …, m, превышает меру его близости с другими классами, то принимается решение о принадлежности этого объекта классу Ω_k.

1.2. Меры близости

В алгоритмах классификации, базирующихся на использовании как детерминированных, так и вероятностных признаков, в качестве меры близости объектов, групп объектов, представляющих собой классы, объекта и группы объектов наиболее часто применяются следующие метрики в виде расстояний.

Рассмотрим ряд расстояний между двумя объектами x_i и x_j.

1. **Евклидово расстояние** $r_1(x_i, x_j) = \sqrt{\sum_{k=1}^{p}(x_i^{(k)} - x_j^{(k)})^2}$.

Евклидово расстояние рекомендуется применять, если признаки $x^{(k)}$, $k = 1, \ldots, p$, однородны по своему физическому смыслу, причем известно, что все они одинаково важны при отнесении объекта к тому или иному классу. Если признаки имеют различную важность, то используется

2. Взвешенное евклидово расстояние $r_2(x_i, x_j) = \sqrt{\sum_{k=1}^{p} w_k (x_i^{(k)} - x_j^{(k)})^2}$,

где w_k – вес k–го признака.

3. **Расстояние Хэмминга** $r_3(x_i, x_j) = \sum_{k=1}^{p} | x_i^{(k)} - x_j^{(k)} |$.

Это расстояние используется в основном как мера различия объектов, признаки которых измерены в шкалах наименований и порядка. Для дихотомических признаков расстояние Хэмминга показывает число несовпадений признаков у анализируемых объектов.

4. **Расстояние Махаланобиса** $r_4(x_i, x_j) = \sqrt{(x_i - x_j)\Sigma^{-1}(x_i - x_j)^T}$,

где Σ^{-1}– матрица, обратная ковариационной матрице. Такое расстояние между объектами удобно тем, что любое линейное преобразование (растяжение, сдвиг, поворот и т. п.) признакового пространства оставляет расстояние Махаланобиса между объектами неизменным. Это расстояние рекомендуется применять, если признаки значительно коррелированы между собой, а также если признаки имеют различную значимость для классификации.

Для характеристики взаимного расположения групп объектов наиболее часто используются следующие виды расстояний.

4. Расстояние между двумя группами S_i и S_j равно расстоянию между ближайшими объектами из этих групп (**расстояние "ближайшего соседа"**):

$r_1(S_i, S_j) = \min \{\rho(x_k, x_m)\}, x_k \in S_i, x_m \in S_j$.

5. Расстояние между двумя группами S_i и S_j равно расстоянию между их математическими ожиданиями: $r_2(S_i, S_j) = \rho(\bar{x}_i, \bar{x}_j)$,

где $\bar{x}_i$ – вектор математических ожиданий для i–й группы.

6. Расстояние между двумя группами S_i и S_j равно расстоянию между самыми дальними объектами этих групп (**расстояние "дальнего соседа"**):

$$r_3(S_i, S_j) = \max \{\rho(x_k, x_m)\}, x_k \in S_i, x_m \in S_j.$$

7. Расстояние между двумя группами S_i и S_j равно среднему арифметическому всевозможных попарных расстояний между объектами рассматриваемых групп: $r_4 \quad (S_i, S_j) = (n_i n_j)^{-1} \sum_{x_k} \sum_{x_m} r(x_k, x_m), \; x_k \in S_i, x_m \in S_j,$

где n_i – число объектов в группе S_i.

В алгоритмах классификации, базирующихся на использовании вероятностных признаков, в качестве меры близости используется ***риск***, связанный с решением о принадлежности распознаваемого объекта к классу Ω_i, i =1, …, m. Пусть даны вероятностные описания классов $(f_i(x), P(\Omega_i))$, $x = (x_1, x_2, ...x_p)$ и риски правильных и ошибочных решений, представляющих собой элементы матрицы потерь вида $C = (c_{ij})$; i, j =1, …, m.

Обозначим a_p событие, состоящее в том, что распознаваемый объект ω описывается набором значений признаков $x_1^0, x_2^0, ...x_p^0$. Тогда значение риска, связанного с решением вида $\omega \in \Omega_k$ при условии, что имеет место событие a_p, выражается как $R(\omega \in \Omega_k | a_p) = R(\Omega_k | a_p) = \sum_{i=1}^{m} c_{ik} \; P(\Omega_i | a_p)$,

где условная апостериорная вероятность $P(\Omega_i | a_p)$ того, что $\omega \in \Omega_k$ в соответствии с теоремой гипотез или формулой Байеса равна

$$P(\Omega_i | a_p) = P(\Omega_i) \, f_i(x_1^0, x_2^0, ...x_p^0) \, / \sum_{i=1}^{m} P(\Omega_i) \, f_i(x_1^0, x_2^0, ...x_p^0).$$

В общем случае решение вида $\omega \in \Omega_k$ принимается, если

$R(\Omega_k | a_p) = \min R(\Omega_i | a_p)$ для всех i =1, …, m.

В алгоритмах распознавания, базирующихся на использовании логических признаков, не используется понятие "мера близости". Когда построено описание классов на языке логических признаков в виде соответствующих булевых соотношений (эквивалентности или импликаций), при подстановке в эти соотношения значений признаков, характеризующих распознаваемый объект,

получаем решение: к какому классу или каким классам объект может быть отнесен и к каким он не относится.

В алгоритмах распознавания, использующих структурные (лингвистические) признаки, понятие меры близости также может не использоваться. Когда построены языки для описания классов в виде совокупности предложений, характеризующих структурные особенности объектов каждого класса, то распознавание неизвестного объекта осуществляется идентификацией предложения, описывающего этот объект, с одним из предложений языка – элемента описания соответствующего класса.

1.3. Виды систем классификации (распознавания)

В зависимости от того, с какого рода информацией работает алгоритм распознавания, системы распознавания (классификации) могут быть разделены на детерминированные, вероятностные, логические, структурные и комбинированные. Каждая из этих систем использует определенные математические методы классификации, реализованные в виде алгоритмов.

Детерминированные системы. В этих системах для построения алгоритмов распознавания используются геометрические меры близости, основанные на измерении расстояний между распознаваемым объектом и эталонами классов. В общем случае применение детерминированных методов распознавания предусматривает наличие координат эталонов классов в пространстве признаков или координат объектов, принадлежащих соответствующим классам.

Вероятностные системы. В таких системах для построения алгоритмов распознавания используются вероятностные методы, основанные на теории статистических решений. В общем случае применение вероятностных методов распознавания предусматривает наличие вероятностных зависимостей между признаками распознаваемых объектов и классами, к которым эти объекты принадлежат.

Логические системы. В этих системах для построения алгоритмов распознавания используются логические методы распознавания, основанные на дискретном анализе и базирующемся на нем исчислении высказываний. В общем случае применение логических методов распознавания предусматривает наличие логических связей, выраженных через систему булевых уравнений, в которой

переменные – логические признаки распознаваемых объектов, а неизвестные величины – классы, к которым эти объекты относятся.

Структурные (лингвистические) системы. В этих системах для построения алгоритмов распознавания используются специальные грамматики, порождающие языки. Языки состоят из предложений, каждое из которых описывает объекты, принадлежащие конкретному классу. Применение структурных методов требует наличия совокупностей предложений для описания множества объектов, принадлежащих всем классам. При этом множество предложений должно быть подразделено на подмножества по числу классов. Элементами подмножеств являются предложения, описывающие объекты, принадлежащие данному подмножеству (классу). Таким образом, априорными описаниями классов являются совокупности предложений, каждое их которых соответствует конкретному объекту, принадлежащему данному классу.

Комбинированные системы. В таких системах для построения алгоритмов распознавания используется специально разработанный метод вычисления оценок. Такие алгоритмы распознавания называют *алгоритмами вычисления оценок* (АВО) [1]. Их применение требует наличия таблиц, где содержатся объекты, принадлежащие соответствующим классам, а также значения признаков, которыми характеризуются объекты. Признаки могут быть детерминированными, логическими, вероятностными и структурными.

Модели для решения задач классификации. В процессе решения прикладных задач обработки данных сформировались семейства (модели) алгоритмов для решения задач классификации. Укажем ряд алгоритмов, получивших практическое применение.

1. Модели, основанные на использовании принципа разделения. Эти модели отличаются главным образом заданием класса поверхностей, среди которых выбирается поверхность (или набор поверхностей), в некотором смысле наилучшим образом разделяющая элементы разных классов [1, 2].

2. Статистические модели. Этот тип моделей алгоритмов распознавания основан на использовании аппарата математической статистики. Они применяются в тех случаях, когда известны или могут быть просто определены вероятностные характеристики классов, например, соответствующие функции распределения [].

3. Модели, построенные на основе так называемого "метода потенциальных функций" [2, 3]. В основе этих моделей лежит заимствованная из физики идея потенциала, определенного для любой точки пространства и зависящего от того, где расположен источник потенциала. В качестве функции принадлежности объекта классу используется потенциальная функция – всюду положительная и монотонно убывающая функция расстояния.

4. Модели вычисления оценок (голосования) [1]. В этих моделях анализируется "близость" между частями описаний ранее классифицированных объектов и объекта, который надо распознать. Наличие близости служит частичным прецедентом и оценивается по некоторому заданному правилу (посредством числовой оценки). По набору оценок близости вырабатывается общая оценка распознаваемого объекта для класса, которая и является значением функции принадлежности объекта классу.

5. Модели, основанные на исчислении высказываний, в частности на аппарате алгебры логики. В этих моделях классы и признаки объектов рассматриваются как логические переменные, а описание классов на языке признаков представляется в форме булевых соотношений.

1. 4. Математическая постановка задачи классификации

Пусть дано множество M объектов ω; на этом множестве существует разбиение на конечное число подмножеств (классов) Ω_i, $i = 1, \ldots, m$, $M = \bigcup_{i=1}^{m} \Omega_i$. Разбиение M определено не полностью. Задана лишь некоторая информация I_0 о классах Ω_i. Объекты ω задаются значениями некоторых признаков x_j, $j = 1, \ldots, p$ (этот набор всегда один и тот же для всех объектов, рассматриваемых при решении определенной задачи). Совокупность значений признаков x_j определяет описание I (ω) объекта ω. Каждый из признаков может принимать значения из различных множеств допустимых значений признаков, например, из следующих: (0, 1) – признак отсутствует или присутствует соответственно; (0, 1, Δ), Δ – информация о признаке отсутствует; (0, 1, …, d – 1) – степень выраженности признака имеет различные градации, $d > 2$; (a_1, …, a_d) – признак имеет конечное число значений, $d > 2$; значениями признака x_j задаются

функции некоторого класса; значениями признака x_j являются функции распределения некоторой случайной величины. Описание объекта $I(\omega) = (x_1(\omega), \ldots, x_n(\omega))$ называют стандартным, если $x_j(\omega)$ принимает значение из множества допустимых значений.

Задача распознавания со стандартной информацией состоит в том, чтобы для данного объекта ω и набора классов Ω_i, $i = 1, \ldots, m$, по обучающей информации $I_0(\Omega_1, \ldots, \Omega_m)$ о классах и описанию $I(\omega)$ вычислить значения предикатов $P_i(\omega) = "\omega \in \Omega_i"$, $i = 1, \ldots, m$. Информация о вхождении объекта ω в класс Ω_i кодируется символами "1" ($\omega \in \Omega_i$), "0" ($\omega \notin \Omega_i$), Δ – неизвестно, принадлежит ли ω классу Ω_i или нет.

1.5. Постановка диагноза как классификационная задача

Постановка диагноза – это классификационная задача. Медицинский диагноз может представлять собой как наименование болезненного состояния человеческого организма, так и наименование той болезненной причины, которая вызвала это состояние. В первом случае речь идет о классификации состояний человеческого организма, которые должны быть диагностированы по некоторым их описаниям, во втором – о классификации причин и диагностике причин, вызвавших изменения состояния организма.

Классификацию заболеваний человека структурно может быть представлена в виде дерева, терминальные вершины которого являются диагнозами. Процесс постановки диагноза при этом представляется как движение по дереву в зависимости от ответов на вопросы, которые ставятся в каждой вершине.

В действительности при постановке диагноза мы лишены возможности получать ответы на прямые вопросы и ограничиваемся косвенными вопросами. Поэтому структура классификации болезней лишь в ограниченной мере может служить целям диагностики и на первый план выдвигается задача формирования диагностических признаков и построения диагностического правила. Эти признаки и диагностическое правило должны быть найдены на основании обучающего статистического материала.

Алгоритмы классификации (распознавания), применяемые в медицинской диагностике, базируются на следующих трех гипотезах. Первая состоит в том, что при полноте описания близким описаниям в пространстве признаков должны

соответствовать близкие диагнозы. Вторая гипотеза выдвигает предположение о функциональном виде разделимости классов: линейная разделимость, квадратичная разделимость и т. п. Третья гипотеза состоит в предположении существования так называемого диагностического сочетания признаков, встречающихся в одном классе значительно чаще, чем в другом.

Задачу диагностирования заболеваний можно поставить следующим образом. Допустим для простоты изложения, что мы имеем дело с двумя заболеваниями D_1 и D_2. Пусть для каждого больного имеется совокупность данных, характеризующих его состояние (анамнез, лабораторные исследования и т. д.) x_1, x_2, …x_p. Будем в дальнейшем называть такие данные признаками (параметрами), а набор признаков x_1, x_2, …x_p – описанием больного.

Задача постановки диагноза состоит в следующем: по имеющемуся набору признаков, полученных у конкретного k–го больного $x_{k1}, x_{k2},..., x_{kp}$, необходимо поставить диагноз D_1 или D_2. Множества признаков для заболеваний D_1 и D_2 могут быть пересекающимися и непересекающимися. В первом случае задача постановки диагноза сводится к построению разделяющей поверхности R (так называемого диагностического правила или дискриминантной функции). Тогда при предъявлении каждого нового больного с целью постановки диагноза определяется, в какую часть пространства от поверхности R попадают значения признаков обследуемого больного. В рассматриваемом случае диагноз может быть поставлен однозначно: у больного заболевание D_1 или D_2.

Гораздо чаще встречается вторая ситуация, когда множества анализируемых признаков для заболеваний D_1 и D_2 являются пересекающимися. Это означает, что хотя бы один из признаков x_j, j =1, …, p, может принимать одно и то же значение для обоих заболеваний. В этом случае задача постановки диагноза также сводится к построению разделяющей поверхности R, но в отличие от первого случая диагноз D_1 или D_2 устанавливается с некоторой вероятностью. Итак, ***задача диагностирования заболеваний математически может быть сведена к построению решающего правила (разделяющей поверхности R) и вычислению вероятности постановки правильного диагноза с помощью этого правила.*** Все сказанное справедливо и для случая, когда нужно диагностировать несколько заболеваний. Тогда диагностическое правило строится для каждой пары заболеваний.

2. МАТЕМАТИЧЕСКИЕ МЕТОДЫ, ИСПОЛЬЗУЕМЫЕ ПРИ ПОСТАНОВКЕ ДИАГНОЗА

В этом разделе рассмотрены математические методы и алгоритмы классификации (распознавания), нашедшие широкое применение в медицинской практике при постановке диагноза.

2.1. Классификация посредством задания границы разделения

Рассмотрим пример построения ***алгоритма распознавания, основанного на принципе разделения.*** Суть этого принципа состоит в следующем. Во многих задачах описания объектов задаются наборами значений числовых признаков (*p* – мерными векторами). Тогда объекты можно интерпретировать как точки *p* – мерного пространства. Их описания, принадлежащие разным классам, могут быть разделены поверхностями достаточно простого вида.

Воспользуемся классом разделяющих поверхностей в виде *гиперплоскостей*

$$\sum_{i=1}^{p} a_i x_i + a_{p+1} = 0.$$

Пусть множество допустимых объектов разделено на два класса: K_1, K_2, $K_1 \cap K_2 = \varnothing$. Пусть также известно, что объекты S_1, …, S_m принадлежат K_1, объекты S_{m+1} , …, S_q – K_2. Эти объекты неравнозначны. Поэтому введем их числовые характеристики $\gamma(S_i) = \gamma_i$ – вес объекта S_i , i = 1, 2,…, m, m +1,…., q. Таким образом, алгоритм распознавания может быть охарактеризован параметрами a_1, …, a_{p+1} – коэффициентами в уравнении гиперплоскости и γ_1, …, γ_q – весами объектов, классификация которых была проведена ранее. Распознавание объекта S_i с описанием $I(S_i)$ производится следующим образом.

Пусть $f(x_1, \ldots, x_p) = \sum_{i=1}^{p} a_i x_i + a_{p+1}$. Разделим объекты S_1, …, S_m на множества K_1^+, K_1^-; $S_i \in K_1^+$, если $f(I(S_i)) > 0$; $S_i \in K_1^-$, если $f(I(S_i)) < 0$. Аналогично объекты S_{m+1} , …, S_q разделим на множества K_2^+, K_2^-. Рассмотрим величины

$$\gamma(K_1^+) = \sum_{S_i \in K_1^+} g(S_i), \quad \gamma(K_1^-) = \sum_{S_i \in K_1^-} g(S_i)$$

и аналогичные им величины $\gamma(K_2^+)$ и $\gamma(K_2^-)$. Вычислим $f(I(S))$. Сопоставим S два числа: $\Gamma_1(S)$, $\Gamma_2(S)$ – соответственно значение функции принадлежности S классам K_1, K_2. Если $f(I(S)) > 0$, то

$$\Gamma_1(S) = (\gamma(K_1^+) + \gamma(K_2^-)) / (\gamma(K_1^-) + \gamma(K_2^+)),$$

$$\Gamma_2(S) = (\gamma(K_2^+) + \gamma(K_1^-)) / (\gamma(K_1^+) + \gamma(K_2^-)).$$

При $f(I(S)) < 0$ $\Gamma_1(S) = (\gamma(K_1^-) + \gamma(K_2^+)) / (\gamma(K_1^+) + \gamma(K_2^-))$ и аналогично вычисляется $\Gamma_2(S)$. По значениям $\Gamma_1(S)$ и $\Gamma_2(S)$ принимается решение об отнесении S к K_1 или K_2. Эта процедура задается **решающим правилом**, которое может быть записано следующим образом:

если $\Gamma_1(S) - \Gamma_2(S) > \delta$, то $S \in K_1$,

если $\Gamma_2(S) - \Gamma_1(S) > \delta$, то $S \in K_2$,

если $|\Gamma_1(S) - \Gamma_2(S)| \le \delta$, то решение не принимается, алгоритм отказывается от классификации S. Здесь δ – параметр решающего правила.

Построенный на принципе разделения алгоритм классификации основан на следующих предположениях: 1) элементы классов K_1 и K_2 разделяются гиперплоскостью; 2) элементы классов не равнозначны по важности, меру этой важности можно выразить числом.

Рассмотрим **примеры применения принципа разделения для диагностики заболеваний.**

Пример 1. Выберем в качестве допустимых решающих правил линейные, а критерием качества – процент ошибок. Для простоты будем считать, что имеются всего два диагноза D_1 и D_2. Будем предполагать, что история болезни описывается p двоичными параметрами x_k, $k = 1,\ldots, p$. Решающее правило можно тогда записать в следующем виде:

$$d(x) = \begin{cases} D_1, \text{ если } \sum_k a_k x_k \ge 0; \\ D_2, \text{ если } \sum_k a_k x_k < 0. \end{cases}$$

Задача заключается в выборе коэффициентов α_k с минимальной ошибкой. Для этого используем метод потенциальных функций [2, 3]. Обозначим значение k–го симптома в i–й истории болезни x_{ik} ($x_{ik} = 1$ или 0). Перенумеруем все истории болезни и будем рассматривать их циклически от №1 до № N и

опять №1 и т. д. Для каждой истории болезни вычисляется $s_i = \sum_k a_k x_{ik}$, где α_k – текущие приближения; в качестве начального приближения можно положить $\alpha_k = 0$. Могут встретиться четыре случая: 1) $s_i \geq 0$ – диагноз D_1; 2) $s_i < 0$ – диагноз D_2; 3) $s_i < 0$ – D_1; 4) $s_i \geq 0$ – D_2. В первых двух случаях ошибки нет. Все α_k остаются без изменения. В третьем случае прибавляется 1 ко всем α_k, для которых $x_{ik} = 1$, остальные α_k не меняются. В четвертом случае из α_k, для которых $x_{ik} = 1$, вычитается 1. Следующее приближение a' связано с предыдущим так:

$$a' = \begin{cases} a & \text{для случая 1 или 2,} \\ a + x_i & \text{для случая 3,} \\ a - x_i & \text{для случая 4.} \end{cases}$$

Пример 2. Предположим, что области двух диагнозов можно разделить плоскостью, так что точки, соответствующие диагнозам D_1 и D_2, лежат по разные стороны от разделяющей плоскости. Для описания с помощью двух признаков уравнение разделяющей плоскости (прямой) имеет вид $f(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$, где x_1 и x_2 – числовые значения признаков (координаты пространства признаков); α_0, α_1 и α_2 – числовые коэффициенты, характеризующие положение разделяющей плоскости.

Решающее правило. Если точка, описывающая признаки больного, находится выше плоскости, то $f(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 > 0$, и устанавливается диагноз D_1. При $f(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 < 0$ устанавливается диагноз D_2. В общем случае для решающего правила составляется линейная функция

$$d(x) = \begin{cases} D_1, \text{ если } \sum_k a_k x_k \geq h; \\ D_2, \text{ если } \sum_k a_k x_k \leq -h; \end{cases}$$

неопределенный ответ, если $-h < \sum_k a_k x_k < h$.

Величины δ_1 и δ_2 и характеризуют порог распознавания. Чем больше значения δ_1 и δ_2, тем выше надежность распознавания, но одновременно тем больше число отказов от установления диагноза.

2.2. Алгоритмы классификации, основанные на вычислении оценок

Принцип действия *алгоритмов классификации, основанных на вычислении оценок (АВО)* состоит в вычислении оценок сходства, характеризующих близость классифицируемого и эталонного объекта (объектов) по подмножеству заданного множества признаков. Важным отличием АВО от других алгоритмов классификации являются более слабые требования к исходной информации, так как они не предусматривают наличия сведений о статистических характеристиках. Исходные данные могут представляться не только в числовой форме, но и задаваться описаниями на естественном языке.

Алгоритм распознавания, основанный на принципе прецедентности или частичной прецедентности, сравнивает описание распознаваемого объекта $I(S)$ с обучающей таблицей $T_{p,m}$ и принимает решение о том, к какому классу отнести объект. Решение выносится на основе вычисления степени сходства распознаваемого объекта (строки таблицы обучения) со строками, принадлежность которых к заданным классам известна. Обучающая таблица $T_{p,m}$ в АВО представляет собой прямоугольную таблицу, строки которой – реализации наблюдений над объектами, столбцы – значения p признаков, характеризующих объект, и индикатор α принадлежности данного объекта одному из заданных классов. Кроме того, в обучающую таблицу включена строка–описание классифицируемого объекта w', в которой отсутствует значение α.

Пусть заданы стандартные описания объектов $\{\omega^i\}$, $\omega^i \in \Omega_i$, и $\{w^{\bar{i}}\}$, $w^{\bar{i}} \notin \Omega_i$. Необходимо определить принадлежность классу Ω_i, $i = 1, \ldots, m$, предъявленного для классификации объекта w'. Если введен способ определения величины близости для некоторых частей описания $J(w')$ и соответствующих частей описаний $J(\omega^i)$ и $J(w^{\bar{i}})$, то можно сформировать характеристику обобщенной близости между объектом w' и множествами объектов ω^i и $w^{\bar{i}}$ соответственно. В простейшем случае обобщенная близость приравнивается к сумме близостей между частями описаний. В результате характеристику вида $\Gamma_i(w') = \Gamma_i^{\Omega_i} - \Gamma_i^{\overline{\Omega}_i}$, где $\Gamma_i^{\Omega_i}$ и $\Gamma_i^{\overline{\Omega}_i}$ – значения соответствующих обобщенных

близостей, можно считать значением функции принадлежности объекта w' классу Ω_i. Величина $\Gamma_i(w')$ называется оценкой объекта w' по классу Ω_i.

Рассмотрим полный набор признаков x_1, … x_p и выделим систему подмножеств множества признаков S_1, …, S_p. Удалим произвольный набор признаков из строк ω_1, ω_2, …, w_{n_m}, w' и обозначим полученные строки $S\tilde{w}_1$, $S\tilde{w}_2$, …, $S\tilde{w}_{n_m}$, $S\tilde{w}'$. Правило близости, позволяющее оценить сходство строк $S\tilde{w}'$ и $S\tilde{w}_r$, состоит в следующем. Пусть «усеченные» строки содержат q первых признаков, т. е. $S\tilde{w}_r = (a_1, \ldots, a_q)$ и $S\tilde{w}' = (b_1, \ldots, b_q)$., и заданы пороги $\varepsilon_1, \ldots\varepsilon_q, \delta$. Строки $S\tilde{w}'$ и $S\tilde{w}_r$ считаются похожими, если выполняется не менее q неравенств вида $|a_j - b_j| \le \varepsilon_{qj}$, $j = 1, \ldots, q$. Величины $\varepsilon_1, \ldots\varepsilon_q, \delta$ являются параметрами класса алгоритмов типа АВО.

Рассмотрим процедуру вычисления оценок по подмножеству S_1. Для остальных подмножеств она аналогична. В обучающей таблице $T_{p,m}$ выделяются столбцы, соответствующие признакам, входящим в S_1; остальные столбцы вычеркиваются. Проверяется близость строки $S\tilde{w}'$ со строками $S\tilde{w}_1$, $S\tilde{w}_2$, …, $S\tilde{w}_{n_1}$, принадлежащими классу Ω_1. Число строк этого класса, близких по выбранному критерию классифицируемой строке $S\tilde{w}'$, обозначается через $\Gamma_{S_1}(w', \Omega_1)$; эта величина представляет собой оценку строки w' для класса Ω_1 по опорному множеству S_1. Аналогично вычисляются оценки для других классов: $\Gamma_{S_1}(w', \Omega_2)$, …, $\Gamma_{S_1}(w', \Omega_m)$.

Применение данной процедуры ко всем остальным опорным множествам алгоритма позволяет получить систему оценок $\Gamma_{S_2}(w', \Omega_1)$, $\Gamma_{S_2}(w', \Omega_2)$, …, $\Gamma_{S_2}(w', \Omega_m)$, …, $\Gamma_{S_l}(w', \Omega_1)$, …, $\Gamma_{S_l}(w', \Omega_m)$. Затем эти оценки суммируются для соответствующих классов по системе опорных множеств алгоритма S_A. На основе анализа суммарных величин принимается решение об отнесении объекта w' к одному из классов Ω_i, $i = 1,\ldots, m$, либо об отказе от его классификации. Решающее правило может принимать различные формы. Классифицируемый объект может быть отнесен к классу, 1) которому соответствует максимальная оценка, 2) для которого оценка будет превышать оценки всех остальных классов не меньше, чем на определенную пороговую величину δ_1, 3) для которого значение соответствующей оценки к сумме оценок для всех остальных классов

будет не менее значения порога δ_2. Аналитические формулы, позволяющие вычислить оценки $\Gamma_i\ (w')$ при различных способах задания системы опорных множеств, приведены в [1].

Последовательная реализация процесса классификации в соответствии с АВО такова: 1) выделяется система опорных множеств алгоритма, по которым производится анализ распознаваемых объектов; 2) вводится понятие близости на множестве частей описаний объектов; 3) задаются правила: а) позволяющие по вычисленным оценкам степени подобия эталонного и распознаваемого объекта вычислить величину, называемую оценкой для пар объектов; б) формирования оценок для каждого из классов по фиксированному опорному множеству на основе оценок для пар объектов; в) формирования суммарных оценок для каждого из классов по всем опорным подмножествам; г) принятия решения, которое на основе оценок для классов обеспечивает отнесение распознаваемого объекта к одному из классов или отказ ему в классификации.

2.3. Статистические методы распознавания (классификации)

В разделах 2.3.1 – 2.3.5. рассматриваются методы классификации для случая, когда распределения вектора x (описания объекта) внутри классов известны. Они задаваются аналитически или с помощью перечисления всех возможных значений x. На основании этой информации строится правило (критерий) классификации.

2.3.1. Критерий отношения правдоподобия как правило классификации

Пусть совокупность объектов разделена на классы Ω_1 и Ω_2, а для характеристики объектов используется признак X. Известны *описания классов* – плотности распределения вероятностей $f_1(x)$ и $f_2(x)$ значений признака x для классов Ω_1 и Ω_2 соответственно (или функции распределения $F_1(x)$ и $F_2(x)$), а также априорные вероятности $P(\Omega_1)$ и $P(\Omega_2)$ появления объектов из классов Ω_1 и Ω_2. В результате эксперимента определено значение x^0 распознаваемого объекта.

Правило отнесения объекта в один из двух заданных классов (правило принятия решения) формулируется следующим образом**: если измеренное**

значение признака у распознаваемого объекта $x^0 > x_0$, то будем относить его в класс Ω_2, а если $x^0 \le x_0$ – в класс Ω_1.

Задача отнесения наблюдения x в один из двух ранее известных классов связана со статистической задачей проверки простой гипотезы против простой альтернативы: H_1: $X \in F_1$ против H_2: $X \in F_2$. Если объект относится к классу Ω_1, а его считают объектом класса Ω_2, то совершается ошибка 1-го рода, условная вероятность которой равна α (по терминологии статистической проверки гипотез это ошибка выбора альтернативной гипотезы H_2 при справедливости H_1). Если справедлива H_2, а выбрана H_1, то совершается ошибка 2-го рода, условная вероятность которой равна β.

При выборе значения x_0 при разделении пространства признака X на два полупространства R_1 и R_2 должны учитываться потери, сопряженные с правильными и ошибочными решениями. Функции потерь, характеризующие потери при совершении ошибок 1-го и 2-го рода, а также потери правильных решений, образуют матрицу потерь вида $C = (c_{ij})$, где c_{11} и c_{22}, c_{12} и c_{21} – потери, связанные соответственно с правильными решениями и ошибками 1-го и 2-го рода. Средний риск в этих условиях равен

$$R(x) = P(\Omega_1)(c_{11}(1-\alpha) + c_{12}\alpha) + P(\Omega_2)(c_{22}(1-\beta) + c_{21}\beta).$$

Значение x_0 выбирается из условия минимума среднего риска:

$$P(\Omega_1)(c_{11} f_1(x_0) - c_{12} f_1(x_0)) + P(\Omega_2)(c_{21} f_2(x_0) - c_{22} f_2(x_0)) = 0,$$

откуда $f_2(x_0) / f_1(x_0) = P(\Omega_1)(c_{12} - c_{11}) / (P(\Omega_2)(c_{21} - c_{22}))$. (1)

Отношение условных плотностей распределения

$$f_2(x) / f_1(x) = l(x) \quad (2)$$

называется отношением правдоподобия. Правая часть (1) (обозначим ее l_0) определяет собой *критическое значение отношения правдоподобия*.

Определим значение x_0 при условии, что x имеет нормальное распределение. Для этого подставим в (2) выражения плотностей распределения $f_1(x)$ и $f_2(x)$ с математическими ожиданиями μ_i и дисперсиями s_i^2, $i = 1, 2$, откуда

$$\exp((x_0-\mu_1)^2 / 2\ s_1^2 - (x_0-\mu_2)^2 / 2 s_2^2) = l_0\ s_1 / s_2. \qquad (3)$$

Решая (3) относительно x_0, получим выражение для вычисления x_0. При $s_1 = s_2 = \sigma$, из (3) находим $x_0 = (\mu_1+\mu_2)/2 + \sigma^2/(\mu_1-\mu_2)\ln l_0$. Если $c_{11} = c_{22} = 0$, $c_{12} = c_{21}$ и $P(\Omega_1) = P(\Omega_2)$, то $x_0 = (\mu_1+\mu_2)/2$.

Значение x_0 позволяет оптимальным образом (в смысле минимума среднего риска) разделить признаковое пространство на области R_1 и R_2. Область R_1 состоит из $x \le x_0$, для которых $l(x) \le l_0$, а область R_2 – из $x > x_0$, для которых $l(x) > l_0$. Поэтому **решение об отнесении объекта к классу Ω_1 следует принимать, если его значение отношения правдоподобия меньше критического значения, решение об отнесении объекта к классу Ω_2 – в противоположной ситуации.**

В общем случае, когда число классов $m > 2$, а объекты описываются p – вектором признаков $x = (x_1, \ldots x_p)$, отношение правдоподобия для классов Ω_k и Ω_l будет $l_{kl} = f_k(x) / f_l(x)$, $k, l = 1, \ldots, m$, матрица потерь вида имеет вид $C = (c_{ij})$, а величина среднего риска равна $R = \sum_{k=1}^{m}\sum_{l=1}^{m} P(\Omega_l) c_{kl} \int_{D_k} f_l(x)dx$.

Из условия минимума среднего риска **уравнение разделяющей границы между областями D_k и D_l, соответствующими классам Ω_k и Ω_l,** будет

$P(\Omega_k) f_k(x) - P(\Omega_l) f_l(x) = 0$ или $\ln(P(\Omega_k) f_k(x) / P(\Omega_l) f_l(x)) = 0$.

Для многомерных нормальных распределений с векторами математических ожиданий μ_k, μ_l и ковариационными матрицами K_k и K_l, $k, l = 1, \ldots, m$, уравнение разделяющей границы между областями D_k и D_l, имеет вид

$$\ln(P(\Omega_k)/P(\Omega_l)) + \ln(f_k(x)/f_l(x)) = \ln(P(\Omega_k)/P(\Omega_l)) - \ln(K_k/K_l) -$$

$$- 0{,}5((x-\mu_k)^{\mathrm{T}} K_k^{-1}(x-\mu_k) - (x-\mu_l)^{\mathrm{T}} K_l^{-1}(x-\mu_l)) = 0.$$

Если $K_k = K_l = K$, то

$$x^{\mathrm{T}} K^{-1} (\mu_k - \mu_l) - (\mu_k - \mu_l)^{\mathrm{T}} K^{-1} (\mu_k - \mu_l) + \ln (P(\Omega_k) / P(\Omega_l)) = 0. \qquad (4)$$

Уравнение (4) – это уравнение гиперплоскости, разделяющей с точки зрения минимальных средних потерь наилучшим образом многомерное пространство признаков на области, соответствующие классам Ω_k и Ω_l, $k, l = 1, \ldots, m$.

Большинство используемых на практике алгоритмов классификации строится на основе (2). В *параметрическом методе* оценивается неизвестный параметр θ предполагаемого теоретического распределения и вместо θ в плотность подставляются оценки $\hat{q}$ и далее вычисляется оценка $l(x)$ в виде

$l(x) = f_2(x, \hat{q}_2) / f_1(x, \hat{q}_1)$.

В *непараметрическом методе* построения алгоритма классификации для данной точки x сразу, минуя оценку параметра $\hat{q}$, строится оценка отношения $f_2(x, \hat{q}_2) / f_1(x, \hat{q}_1)$.

2.3.2. Критерий Байеса

Критерий Байеса – это правило, в соответствии с которым решение об отнесении объекта x в один из двух ранее известных классов выбирается таким образом, чтобы обеспечить минимальную вероятность ошибочной классификации (минимум среднего риска). Байесовское решающее правило является частным случаем критерия отношения правдоподобия.

Минимум риска, усредненный по множеству решений задачи классификации, обеспечивается тогда, когда ***решение о принадлежности объекта к классу*** Ω_1 ***или*** Ω_2 ***принимается в соответствии со следующим правилом:*** **если измеренное значение признака у данного объекта расположено в области R_1, то объект относится к классу Ω_1, если в области R_2 – к классу Ω_2.**

Стратегия, основанная на этом правиле, называется *байесовской стратегией*, а минимальный средний риск – *байесовским риском*. Байесовская стратегия может быть описана следующим образом. Пусть в результате эксперимента установлено, что значение признака у распознаваемого объекта ω составляет $x = x^0$. Тогда условная вероятность принадлежности объекта к классу Ω_1 (условная вероятность первой гипотезы в соответствии с теоремой гипотез или формулой Байеса) равна $P(\Omega_1 | x^0) = P(\Omega_1) f_1(x^0) / f(x^0)$, а условная

вероятность принадлежности объекта к классу Ω_2 (условная вероятность второй гипотезы) – $P(\Omega_2 | x^0) = P(\Omega_2) f_2(x^0) / f(x^0)$, где $f(x^0) = P(\Omega_1) f_1(x^0) + P(\Omega_2) f_2(x^0)$, а $P(\Omega_i | x^0)$ – апостериорная вероятность принадлежности распознаваемого объекта классу Ω_i, $i = 1, 2$.

Условные риски, связанные с решениями $\omega \in \Omega_1$ и $\omega \in \Omega_2$, равны соответственно

$$R(\Omega_1 | x^0) = c_2 P(\Omega_2 | x^0), \quad R(\Omega_2 | x^0) = c_1 P(\Omega_1 | x^0). \qquad (5)$$

Байесовская стратегия решает задачу с минимальным условным риском. Это значит, что предпочтение решению $\omega \in \Omega_1$ следует отдавать тогда и только тогда, когда $R(\Omega_1 | x^0) / R(\Omega_2 | x^0) < 1$. Подставим в это выражение значения условных рисков, определенных в (5). Тогда неравенство $c_1 P(\Omega_1 | x^0) > c_2 P(\Omega_2 | x^0)$ или $P(\Omega_1 | x^0) / P(\Omega_2 | x^0) > c_2 / c_1$ определяет, в каких условиях необходимо принять решение о том, что $\omega \in \Omega_1$.

Рассмотрим другую форму записи байесовского критерия отнесения объекта к соответствующему классу. Пусть имеются классы Ω_1 и Ω_2. Априорные вероятности появления объектов этих классов равны соответственно p_1 и p_2, $c_{11} = c_{22} = 0$, $c_{12} = c_1$, $c_{21} = c_2$. Известны многомерные плотности распределения вероятностей значений признаков $f_1(x_1^0, \ldots, x_p^0)$ и $f_2(x_1^0, \ldots, x_p^0)$ для первого и второго классов. Выразим средний риск R через ошибки первого α и β второго рода: $R = c_1 p_1 \alpha + c_2 p_2 \beta$, а ошибки первого α и β второго рода через p–мерные интегралы от плотностей вероятностей по областям R_1 и R_2. Задача состоит в минимизации среднего риска. Это достигается при условии, что $c_2 p_2 f_2(x_1, \ldots, x_p) - c_1 p_1 f_1(x_1, \ldots, x_p) < 0$. Отсюда следует известное **решающее правило: распознаваемый объект ω, признаки которого равны x_j^0, $j = 1, \ldots, p$, относится к классу Ω_1, если** $f_2(x_1^0, \ldots, x_p^0) / f_1(x_1^0, \ldots, x_p^0) > c_1 p_1 / c_2 p_2$, где $c_1 p_1 / c_2 p_2 = l_0$ – пороговое значение отношения правдоподобия. **В противном случае объект ω относится к классу Ω_2.**

Использование байесовского правила в медицинской диагностике. В клинической практике известны симптомы, наличие которых однозначно определяет заболевание. С другой стороны встречаются симптомы, исключающие тот или иной диагноз. Чаще всего основные, определяющие клинику симптомы отсутствуют и любой из них может встречаться с некоторой частотой при различных заболеваниях. Естественным поэтому является использование вероятностных методов для постановки диагноза [3, 4].

Пусть необходимо проводить дифференциальную диагностику между заболеваниями D_1, D_2,..., D_m. Для каждого из них характерно распределение условных вероятностей $P(S \mid D_j)$ появления у больного того или иного симптомокомплекса $S = (S_1,..., S_i,..., S_p)$, где S_i – возможные значения (градации) различных симптомов (i = 1, …, p). Если эти распределения, а также априорные вероятности заболеваний $P(D_j)$ заданы, то задача дифференциальной диагностики сводится к статистической задаче выбора гипотез, оптимальное диагностическое правило для которой можно построить с помощью критерия Байеса. Рассмотрим два способа применения этого критерия.

1–й способ. Воспользуемся вычислением среднего условного риска в соответствии с формулой (5) для числа классов m = 2. Запишем (5) в виде

$l_{12}(S) = P(S \mid D_1) / P(S \mid D_2) > l_0$.

Тогда **диагностическое правило** можно сформулировать следующим образом:

установить больному диагноз D_1, если $l_{12}(S) > l_0$;

установить больному диагноз D_2, если $l_{12}(S) < l_0$.

При $m > 2$ диагностическое правило **имеет вид: больному, характеризуемому набором симптомов S, следует поставить диагноз D_i, если $R_i(S) < R_j(S)$, т. е. если**

$$\sum_{k=1}^{m} c_{ki} P(S \mid D_k) P(D_k) < \sum_{q=1}^{m} c_{qj} P(S \mid D_q) P(D_q),\ i, j = 1,\ldots,m,\ i \neq j.$$

2–й способ. Вычислим апостериорную вероятность диагноза D_j по формуле Байеса:

$$P(D_j \mid S_i) = \frac{P(D_j) P(S \mid D_j)}{\sum_j P(D_j) P(S \mid D_j)}, \qquad (6)$$

где $P(D_j)$ – априорная вероятность заболевания с диагнозом D_j среди рассматриваемой группы болезней; $P(S \mid D_j)$ – вероятность появления комплекса признаков при диагнозе D_j.

Решающее правило формулируется следующим образом. **Больному приписывается заболевание D_k, для которого вероятность $P(D_k \mid S)$ значительно превосходит вероятности $P(D_j \mid S)$ для других заболеваний**

(j^1 k).

Трудность применения такого решающего правила состоит в том, что распределения $P(S \mid D_j)$ и вероятности $P(D_j)$ не заданы. Вероятность заболевания $P(D_j)$ можно оценить, вычислив на достаточно большом клиническом материале частоты, с которыми рассматриваемые болезни встречаются. Оценку распределения $P(S \mid D_j)$ сделать значительно сложнее. Это связано с тем, что для каждого диагноза необходимо определить условную вероятность любой комбинации признаков. Но уже для 30 двоичных клинических параметров мы имеем свыше миллиарда симптомокомплексов, составленных из принятых признаков. Собрать клинический массив для оценки условных вероятностей $P(S \mid D_j)$ невозможно. Единственным выходом в данной ситуации является использование вместо $P(S \mid D_j)$ какой-нибудь аппроксимации, для оценки и запоминания которой не требуется столь больших ресурсов. Способ аппроксимации распределения $P(S \mid D_j)$ определяется вводимыми нами предположениями.

Случай независимых признаков. Наиболее распространенным является предположение, что события, состоящие в появлении у больных тех или иных значений рассматриваемых нами симптомов, статистически независимы. Тогда в (6) можно выразить $P(S \mid D_j)$ через условные вероятности появления отдельных значений симптомов при заболевании D_j: $P(S \mid D_j) = \prod_i P(S_i \mid D_j)$.

Теперь для вычисления $P(S \mid D_j)$ достаточно иметь Kn чисел $P(S_i \mid D_j)$ и n чисел $P(D_j)$. Для десятков заболеваний и нескольких сотен признаков объем исходной информации не превысит нескольких десятков тысяч чисел, что вполне доступно для современных ЭВМ. Кроме того, $P(S_i \mid D_j)$ легко оценить, подсчитав частоту появления значений симптома S_i при заболевании D_j на достаточно большом количестве историй болезни.

Рассмотрим этот широко распространенный алгоритм, так как он позволяет получить хорошие результаты.

Медицинские сведения используются в виде диагностической таблицы, содержащей вероятности появления признаков для данной группы заболеваний. Представим для простоты, что диагностическая таблица составлена для трех

заболеваний и содержит четыре признака. Введем следующие обозначения. Диагнозы: D_1 – тетрада Фалло; D_2 – дефект межпредсердной перегородки; D_3 – незараще нный артериальный проток; признаки: S_1 – цианоз; S_2 – усиление легочного рисунка; S_3 – акцент 2-го тона во втором межреберье слева; S_4 – правограмма (ЭКГ).

Признак S_p будет достоверным, если при данном заболевании он встречается в 100% случаев (например, сыпь при кори, белок в моче при остром нефрите и т. д.). Вероятность такого признака принимается за единицу. В общем случае вероятность признака при заболевании D_j: $P(S_i \mid D_j) = n_{ij} / (n_j - n_{i0})$,где n_{ij} – число больных с диагнозом D_j , имеющих признак S_i; n_j – общее число больных с данным заболеванием; n_{i0} – число больных с диагнозом D_j, не обследованных на наличие признака S_i.

Величина $P(S_i \mid D_j)$ определяется на основании данных медицинской статистики, результатов обработки архивного материала и литературных данных, причем достоверность $P(S_i \mid D_j)$ будет тем выше, чем больше n.

Предположим, что по данным статистики при D_1 признак S_1 встречается в 90 % случаев, S_2 не встречается (0 %), S_3 встречается в 5 % случаев, S_4 встречается в 60 % случаев; при D_2 признак S_1 встречается в 15 %, S_2, S_3, S_4 – в 80 % случаев; при D_3 признак S_1 встречается в 100 %, S_2 – в 95 %, S_3 – в 90 %, S_4 – в 10 % случаев. Будем считать также, что априорная вероятность заболевания D_1 равна 35 %, D_2 – 15 %, D_3 –50 %.

В диагностической таблице эти данные могут быть представлены следующим образом.

Таблица 1

Диагноз	$P(D_j)$	$P(S_1 \mid D_j)$	$P(S_2 \mid D_j)$	$P(S_3 \mid D_j)$	$P(S_4 \mid D_j)$
D_1	0,35	0,90	0	0,05	0,60
D_2	0,15	0,15	0,80	0,80	0,80
D_3	0,50	0,10	0,95	0,90	0,10

Апостериорная вероятность диагноза D_j при наличии комплекса признаков S определяется по приведенной выше формуле Байеса (6).

Предполагая признаки независимыми, будем иметь в рассматриваемом

случае

$$P(S\,|D_j) = P(S_1\,|D_j)\,P(S_2\,|D_j)\,P(S_3\,|D_j)\,P(S_4\,|D_j); \qquad (7)$$

$$P(S) = \sum_{j=1}^{3} P(D_j)\,P(S_1\,|D_j)\,P(S_2\,|D_j)\,P(S_3\,|D_j)\,P(S_4\,|D_j). \qquad (8)$$

Вычислим вероятности диагнозов сначала в том случае, когда у больного проявились все четыре признака. Тогда из формул (6) – (8) находим вероятности диагнозов $P(D_1 \mid S) = 0$; $P(D_2 \mid S) = 0{,}73$; $P(D_3 \mid S) = 0{,}27$.

В диагностике по методу Байеса предполагается, что у больного имеется одно из заболеваний, содержащихся в диагностической таблице. Таким образом, наиболее вероятным оказывается диагноз D_2. Теперь рассмотрим случай, когда у больного отсутствует признак S_1 (цианоз), но имеются все остальные признаки. Вероятность отсутствия признака S_1 равна $P(\bar{S}_1 \mid D_j) = 1 - P(S_1 \mid D_j)$.

Расчет производится точно таким же образом, но вероятность $P(S_1 \mid D_j)$ в формуле (6) заменяется на $1 - P(S_1 \mid D_j)$. В результате получим $P(D_1 \mid \bar{S}_1) = 0$; $P(D_2 \mid \bar{S}_1) = 0{,}63$; $P(D_3 \mid \bar{S}_1) = 0{,}37$.

В табл.2 приведены вероятности диагнозов при наличии всех признаков и при отсутствии одного из них.

Таблица 2

Диагноз	**Признаки**				
	S_1, S_2, S_3, S_4	$\bar{S}_1, S_2, S_3, S_4$	$S_1, \bar{S}_2, S_3, S_4$	$S_1, S_2, \bar{S}_3, S_4$	$S_1, S_2, S_3, \bar{S}_4$
D_1	0	0	0,75	0	0
D_2	0,73	0,63	0,23	0,86	0,07
D_3	0,27	0,37	0,02	0,14	0,93

Из табл. 2 видно, что наиболее четкая картина (по отношению к вероятностям заболеваний) в рассматриваемом примере получается в том случае, когда у больного отсутствует признак S_4, но имеются все остальные признаки: $P(D_3) = 0{,}93$.

Полученные вероятности сравниваются с некоторой пороговой величиной T_j. **Решающее правило формулируется следующим образом. Если $P(D_j) > T_j$, то делается вывод о наличии у больного диагноза D_j (обычно $T_j > 0{,}9$).**

Случай зависимых признаков. Пусть у больного, характеризуемого набором симптомов $S = (S_1, \ldots, S_p)$, необходимо диагностировать одно из

возможных заболеваний D_j, $j = 1,\ldots, m$. Предположим, что события, состоящие в появлении у больного тех или иных значений рассматриваемых симптомов, статистически зависимы. В этом случае вероятность того, что у больного с симптомокомплексом S может быть заболевание D_j, выражается по формуле Байеса (6) следующим образом:

$$P(D_j|\, S_1, \ldots, S_p) = \frac{P(D_j)P(S_1 \mid D_j)P(S_2 \mid D_j, S_1)\ldots P(S_p \mid D_j, S_1, S_2, \ldots, S_{p-1})}{\sum_j P(D_j)P(S_1 \mid D_j)P(S_2 \mid D_j, S_1)\ldots P(S_p \mid D_j, S_1, S_2, \ldots, S_{p-1})}, \quad (9)$$

где $P(S_i \mid D_j, S_1, S_2, \ldots, S_{i-1})$ – условная вероятность появления признака S_i при наличии признаков $S_1, \ldots, S_{i-1}$ для диагноза D_j, ($i = 1, \ldots, p$; $j = 1, \ldots, m$).

Диагноз ставится на основании **решающего правила: больному приписывается заболевание D_k, для которого вероятность $P(D_k \mid S_1, \ldots, S_p)$ значительно превосходит вероятности $P(D_j \mid S_1, \ldots, S_p)$ для других заболеваний ($j \neq k$).**

2.3.3. Минимаксный критерий

При построении систем классификации (распознавания) возможна ситуация, когда априорные вероятности появления объектов соответствующих классов неизвестны. Поэтому минимизировать значение среднего риска принятия решения на основании байесовской стратегии не представляется возможным. В этой ситуации используется критерий, который минимизирует максимально возможное значение среднего риска. Этот критерий называется ***минимаксным критерием***.

Минимаксная стратегия состоит в том, что решение о принадлежности неизвестного объекта определенному классу принимается на основании байесовской стратегии, соответствующей такому значению p_1, при котором средний риск максимален.

Пусть значения $P(\Omega_i)$, $i = 1, \ldots, m$, неизвестны. При наличии двух классов Ω_1 и Ω_2 байесовский риск с учетом того, что $P(\Omega_1) = p_1$, $P(\Omega_2) = 1 - p_1$, $c_{11} = c_{22} = 0$, $c_{12} = c_1$, $c_{21} = c_2$, равен $R_{\min} = c_1 p_1 \alpha + c_2 (1 - p_1) \beta$. $R_{\min}$ является функцией p_1: $R_{\min} = R(p_1)$, которая равно нулю при $p_1 = 0$ и $p_1 = 1$. Пусть $R_{\min}$ достигает своего наибольшего значения при $p_1 = p_1^0$. Этот риск представляет собой максимальное значение минимального байесовского риска. Применение минимаксного критерия означает, что при отсутствии данных относительно априорных вероятностей появления объектов следует ориентироваться на $p_1 = p_1^0$. Средние потери при $p_1 = p_1^0$ определяются как $R = c_1 p_1 \alpha^0 + c_2 (1 - p_1) \beta^0$,

где $\alpha^0 = \alpha\ (p_1^0)$, $\beta^0 = \beta\ (p_1^0)$ – ошибки первого и второго рода при априорной вероятности $p_1 = p_1^0$. Минимаксная стратегия обеспечивает, что при $p_1 < p_1^0$ и $p_1 > p_1^0$ средние потери не будут превышать максимального значения минимальных средних (байесовских) потерь.

2.3.4. Процедура последовательных решений (метод Вальда)

Ранее предполагалось, что решение о принадлежности классифицируемого объекта ω соответствующему классу Ω_i, $i = 1, \ldots, m$, принимается после измерения всех признаков объекта x_j, $j = 1, \ldots, p$. Однако возможен другой подход к решению этой задачи: после измерения каждого очередного признака x_1; x_1, x_2; x_1, x_2, x_3 и т. д. решается задача классификации на основании измеренных к текущему моменту признаков неизвестного объекта. При этом в зависимости от результатов сравнения полученного решения с некоторой установленной заранее границей либо измеряется очередной признак объекта, либо прекращается накопление информации об этом объекте. Такая процедура решения задачи распознавания называется последовательной.

Метод Вальда широко используется для решения задач дифференциальной диагностики. Он представляет собой последовательную процедуру обследований, при которой достигается выбранный уровень вероятности диагноза. Сущность метода состоит в следующем.

Пусть больной характеризуется набором признаков $X = (x_1, \ldots, x_p)$. Предположим, что требуется установить один из двух возможных диагнозов D_1 и D_2. Сначала проводится обследование по признаку x_1. Предположим далее, что при диагнозе D_1 признак x_1 имеет частоту встречаемости $P(x_1|D_1)$, при диагнозе D_2 – соответственно $P(x_1|D_2)$. Если у больного отмечается наличие признака x_1, а при диагнозе D_1 он встречается значительно чаще, чем при D_2, то можно сделать вывод в пользу диагноза D_1.

Решающее правило:

принимается диагноз D_1, если

$$P(x_1|D_1) / P(x_1|D_2) > A, \qquad (10)$$

где A – верхняя граница, необходимая для принятия решения.

В противоположном случае, когда признак x_1 значительно чаще встречается при диагнозе D_2, принимается решение в пользу диагноза D_2, если

$$P(x_1|D_1) / P(x_1|D_2) < B, \quad (11)$$

где B – нижняя граница отношения.

Если отношение вероятностей, называемое отношением правдоподобия, удовлетворяет неравенству

$$B<P(x_1|D_1) / P(x_1|D_2) < A, \quad (12)$$

то **требуется провести дополнительное обследование по признаку x_2** и т. д.

Эта процедура основана на теореме Байеса и может быть получена следующим образом. Рассмотрим общий случай, когда признаки $x_1, \ldots, x_p$ статистически зависимы. Вероятность $P(D_1|x_1,\ldots,x_p)$ того, что у больного с набором признаков $X = (x_1, \ldots, x_p)$ может быть заболевание D_1, выражается по формуле (6). Аналогично вычисляется условная вероятность $P(D_2|x_1,\ldots,x_p)$ возникновения заболевания D_2. Найдем отношение

$$P(D_1|x_1,\ldots,x_p) / P(D_2|x_1,\ldots,x_p) = (P(D_1) / P(D_2)) \cdot (P(x_1|D_1) / P(x_1|D_2)) \cdot \ldots$$

$$\cdot \ldots \cdot (P(x_p|D_1, x_1,\ldots,x_{p-1}) / P(x_p|D_2, x_1,\ldots,x_{p-1})). \quad (13)$$

Логарифм каждого множителя в правой части (13) называется диагностическим коэффициентом. Процедуру диагностирования на основе (13) можно организовать следующим образом.

В процессе постановки диагноза возможны ошибки двух видов. Ошибка первого вида (ее вероятность α) совершается, когда больному с диагнозом D_1 устанавливается диагноз D_2. Ошибка второго вида (ее вероятность β) состоит в том, что больному с диагнозом D_2 устанавливается диагноз D_1. Опасность этих ошибок может быть неодинаковой. Будем считать пороговым такое превышение вероятности одной из диагностических гипотез "заболевание D_1" или "заболевание D_2", которое соответствует требуемому превышению частоты правильных диагнозов над частотой ошибочных. Обозначим $P(D_1^+)$ – вероятность установления больному с заболеванием D_1 правильного диагноза; $P(D_2^+)$ – вероятность установления правильного диагноза больному с заболеванием D_2; $P(D_1^-)$ – вероятность установления больному с

заболеванием D_2 ошибочного диагноза D_1; $P(D_2^-)$ – вероятность установления больному с заболеванием D_1 диагноза D_2.

Выберем пороговую величину для принятия решения. Ее можно задать в соответствии с принятым допустимым отношением правильных и ошибочных диагнозов как $B = P(D_1^+) / P(D_1^-)$. Учитывая, что $P(D_1^+) = 1 - \alpha$ и $P(D_1^-) = \beta$, то $B = (1 - \alpha) / \beta$ – порог для выбора гипотезы "заболевание D_1". Аналогично устанавливается порог для второго заболевания $A = P(D_2^-) / P(D_2^+) = \alpha / (1 - \beta)$.

Процесс принятия решения при последовательной диагностической процедуре сводится к следующему. **Если**

$$P(D_1 | x_1,\ldots,x_p) / P(D_2 | x_1,\ldots,x_p, D_1) \geq B = (1 - \alpha) / \beta \text{ или}$$

$$P(D_2 | x_1,\ldots,x_p) / P(D_1 | x_1,\ldots,x_p, D_2) \leq A = \alpha / (1 - \beta),$$

то принимается соответственно решение "заболевание D_1" или "заболевание D_2". Если

$$\alpha / (1 - \beta) < P(D_1 | x_1,\ldots,x_p) / P(D_2 | x_1,\ldots,x_p, D_1) < (1 - \alpha) / \beta,$$

то необходимо ввести в рассмотрение новые данные о больном и продолжить диагностическую процедуру. Это соответствует введению новых диагностических коэффициентов. Чтобы облегчить вычисления, выражение (13) логарифмируется. Таким образом, **на каждом шаге диагностического процесса суммируются логарифмы диагностических коэффициентов и полученная сумма сравнивается с порогами. Если хотя бы один порог достигается, то диагностический процесс прекращается и принимается соответствующее решение.**

Если число классов $m > 2$, то процедура последовательного анализа состоит в следующем. Зададим допустимое значение вероятности правильного e_{ii} и ошибочного e_{iq} решения, что позволит определить значения порога для каждого класса, т. е. $A(\Omega_i) = (1 - e_{ii}) / (\prod_{i=1}^{m}(1 - e_{iq}))^{m-1}$, $i = 1, \ldots, m$, $q \neq i$. Пусть определен вектор признаков $\mathbf{x}_s^0 = (x_1^0, \ldots, x_s^0)$ распознаваемого объекта и вычислено отношение вероятностей для каждого класса:

$$l_s(\mathbf{x}_s^0 | \Omega_i) = f_i(\mathbf{x}_s^0) / (\prod_{i=1}^{m} f_i(\mathbf{x}_s^0))^{m-1}.$$

Сопоставим $l_s(\mathbf{x}_s^0 | \Omega_i)$ с порогом $A(\Omega_i)$, $i = 1, \ldots, m$. Если $l_s(\mathbf{x}_s^0 | \Omega_i) < A(\Omega_i)$, то принимается решение о том, что $\omega \in \Omega_i$. Если апостериорная информация о признаках объекта не позволяет исключить все классы, кроме одного, то проводится следующий эксперимент с целью определения x_{s+1}. После этого определяется $l_{s+1}(\mathbf{x}_{s+1}^0 | \Omega_i)$ и производится его сравнение с порогом и т. д.

2.3.5. Дискриминантный анализ

В данном разделе изложен метод построения правила (критерия) классификации в условиях, когда распределения признака x внутри классов определены лишь частично. При этом используются два вида информации: 1) предположения о свойствах распределений (гладкость, принадлежность плотности распределения $f_j(x)$, $j = 1,\ldots,m$ к некоторому известному параметрическому классу) и 2) обучающая выборка.

Дискриминантным анализом **(ДА) называется метод, с помощью которого на основании обучающей выборки и предположений строится конкретное правило классификации.**

Задача дискриминантного анализа. В общем виде задача различения (дискриминации) формулируется следующим образом. Пусть результатом наблюдения над объектом является реализация p – мерного случайного вектора $x = (x_1, \ldots x_p)^T$. Требуется построить правило отнесения наблюдения x к одной из возможных совокупностей (классов) Ω_i, $i = 1, \ldots, m$.

Для построения правила классификации (решающего правила – РП) все выборочное пространство R значений вектора $x \in R$ разбивается на области R_i ($i = 1, \ldots, m$) так, что при попадании x в R_i объект относят к классу Ω_i. РП выбирается в соответствии с определенным принципом оптимизации на основе априорной информации о совокупностях извлечения объекта из Ω_i. Априорная информация может быть представлена в виде: 1) некоторых сведений о функции p – мерного распределения признака в совокупности, 2) в виде выборок их этих совокупностей. Рассмотрим РП для двух указанных случаев.

1. Классификация в случае, когда распределения классов определены полностью. Модель двух нормальных распределений с общей ковариационной матрицей (модель Фишера) [2].

Предположения. Теоретические распределения в данном случае являются многомерными нормальными $N(M_1, \mathbf{\Sigma})$ и $N(M_2, \mathbf{\Sigma})$, M_1 и M_2 – векторы математических ожиданий первого и второго класса, $\mathbf{\Sigma}$ – общая ковариационная матрица, $|\mathbf{\Sigma}| > 0$.

Правило классификации определяется с помощью неравенств

$$h(x) = (x - (M_1 + M_2)/2)^{\mathrm{T}} \ \mathbf{\Sigma^{-1}} (M_2 - M_1) \quad c$$

и формулируется следующим образом: *если значение функции классификации $h(x)$ для распознаваемого объекта с описанием x меньше некоторого порогового значения c, то объект относят в первый класс, в противном случае – во второй.*

Разделяющая граница представляет собой гиперплоскость в p – мерном пространстве, касательную в одной и той же точке к одной из линий постоянного уровня плотности $N(M_1, \mathbf{\Sigma})$ и одной из линий постоянного уровня плотности $N(M_2, \mathbf{\Sigma})$.

Модель двух нормальных распределений с разными ковариационными матрицами.

Предположения. Теоретические распределения в этом случае $N(M_i, \mathbf{\Sigma_i})$, $|\mathbf{\Sigma_i}| > 0$, $i = 1, 2$.

Правило классификации определяется с помощью неравенств

$$h(x) = (x - M_1)^{\mathrm{T}} \ \Sigma_1^{-1} (x - M_1) - (x - M_2)^{\mathrm{T}} \ \Sigma_2^{-1} (x - M_2) + \ln(|\mathbf{\Sigma_1}| / |\mathbf{\Sigma_2}|) \quad c$$

и формулируется аналогично случаю с общей ковариационной матрицей.

Разделяющая граница $h(x)$ представляет собой полином второго порядка от x.

Метод получения вида разделяющей поверхности для двух рассмотренных моделей основан на методе отношения правдоподобия (см. 2.3.1.), где константа c известна, если известны априорные распределения и потери от неправильной классификации, или неизвестна, если отсутствует информация о потерях и априорных вероятностях.

Классификация при наличии обучающих выборок

Выборка **представляет собой последовательность независимых пар наблюдений вида $W_n = \{(X_j, y_j),\ j = 1, \ldots, n\}$.** Здесь y_j показывает номер класса, которому принадлежит наблюдение X_j; $P\{y_j = i\} = p_i$, $i = 1, \ldots, m$; p_i – неизвестная вероятность того, что X будет извлечено из i – го класса; число

классов m известно; $\sum_{i=1}^{m} p_i = 1$. Для $y_j = i$ все X_j распределены с неизвестной функцией распределения F_i. Число $y_j = i$ в выборке обозначается n_i и называется объемом выборки из i – го класса.

Предположение. F_i принадлежит некоторому известному семейству распределений, зависящему от неизвестного параметра θ_i, i =1, …, m;. Этот случай называется *параметрическим.*

В параметрическом случае вид разделяющей поверхности известен с точностью до параметров, зависящих от θ_i. В ДА наиболее часто используются так называемые ***подстановочные алгоритмы*** построения правила классификации**,** в которых неизвестные в отношении правдоподобия параметры модели заменяются их оценками, построенными по обучающей выборке, а затем определяются параметры разделяющих функций. Рассмотрим пример такого алгоритма, использующего часть информации предположений и выборку.

Подстановочный алгоритм в задаче Фишера [2]

Предположения: теоретические распределения $F_i = N(M_i, \mathbf{\Sigma})$, $|\mathbf{\Sigma}| > 0$, i =1, 2; параметры распределения: M_i (векторы математических ожиданий), $\mathbf{\Sigma}$ (ковариационная матрица) – неизвестны, матрица $\mathbf{\Sigma}$ – общая для обоих распределений.

M_i и $\mathbf{\Sigma}$ могут быть оценены по выборочным данным. Заданы две выборки $x_i = (x_1^{(i)}, ..., x_{n_i}^{(i)})$, i =1, 2. Оценками $\bar{X}_q$ (q = 1, …, p) компонентов вектора математических ожиданий M_q являются

$$\bar{X}_q^{(i)} = 1/n_i \sum_{j=1}^{n_i} x_{jq}^{(i)}, \; q = 1, \ldots, p. \qquad (14)$$

Оценки $(S_i)_{rs}$ элементов ковариационной матрицы $\mathbf{\Sigma}$:

$$(S_i)_{rs} = 1/(n_i - 1) \sum_{j=1}^{n_i} (x_{jr}^{(i)} - \bar{X}_r^{(i)})(x_{js}^{(i)} - \bar{X}_s^{(i)}); \; r, s = 1, \ldots, p. \qquad (15)$$

Получив оценки (14) и (15), подставим их в РП, построенное на основе отношения правдоподобия:

$$f(x, \bar{X}^{(2)}, \boldsymbol{S}) / f(x, \bar{X}^{(1)}, \boldsymbol{S}) \quad c, \qquad (16)$$

где $\boldsymbol{S} = 1/(n_1 + n_2 - 2)\sum_{i=1}^{2}(n_i - 1)\sum_{j=1}^{n_i}(x_j^{(i)} - \bar{X}^{(i)})(x_j^{(i)} - \bar{X}^{(i)})^T$,

$f(x, \bar{X}, \boldsymbol{S}) = (2p)^{-p/2} |S|^{-1/2} \exp(-(X - \bar{X})^T S^{-1}(X - \bar{X})/2)$ – плотность p – мерного нормального распределения.

Правило классификации формулируется следующим образом: ***если вычисленное значение отношения правдоподобия*** **(16)** ***больше некоторого заданного порогового значения, то классифицируемый объект принадлежит второму классу, в противном случае – первому классу.***

При числе классов $m > 2$ класс, которому принадлежит объект x, можно определить на основе неравенств $f_i(x) > f_j(x)$; $i, j = 1, \ldots, m$, $i \neq j$.

Для случая, когда $f_i(x)$ – плотности многомерного нормального распределения, путем преобразования выражения (16) получим **уравнение разделяющей границы** в виде $h_i(x) = (x, n^{(i)} - g^{(i)})$, где $n^{(i)} = \boldsymbol{S}^{-1} \bar{X}^{(i)}$, $g^{(i)} = \frac{1}{2} (\bar{X}^{(i)})^T \boldsymbol{S}^{-1} \bar{X}^{(i)}$, $i = 1, \ldots, m$.

Правило классификации: распознаваемый объект x относится к классу $i = 1, \ldots, m$, для которого значение линейной функции $h_i(x)$ является максимальным.

2.3.6. Характеристики качества классификации

Для характеристики простого правила классификации при двух классах в условиях полностью известных распределений используются ошибки первого α и второго β рода, а также вероятность того, что наблюдение извлечено из одного из классов. В общем случае статистический критерий классификации может быть представлен в виде $\gamma(x)$ c, где γ – известная функция x, а c – порог критерия.

Назовем объекты первой совокупности (первого класса) "случаями" (случай заболевания, случай дефекта и т. п.), а объекты второй совокупности (второго класса) – "не–случаями". Пусть далее принимается гипотеза, что объект с описанием x является случаем, если $\gamma(x) < c$, и гипотеза, что объект является не–случаем, если $\gamma(x) \geq c$.

Результаты классификации удобно представить в виде таблицы [2].

Таблица 3

Результаты применения критерия	**Статус объекта**		**Всего**
	"случай"	**"не–случай"**	
Принимается гипотеза "случай" $\gamma(x) < c$	a	b	$a + b$
Отвергается гипотеза "случай" $\gamma(x) \geq c$	e	f	$e + f$
Всего	$a + e$	$b + f$	n

В медицинской практике используются следующие характеристики, получаемые с помощью таблицы .

Частота случаев: $(a + e) / n$.

Чувствительность критерия в предсказании случая $a / (a + e)$, т. е. доля случаев, для которых $\gamma(x) < c$. Чувствительность может быть выражена через ошибку первого рода α как $1 - \alpha$.

Специфичность критерия: $f / (b + f)$, т. е. доля не–случаев, для которых $\gamma(x) \geq c$. Специфичность равна $1 - \beta$, где β – ошибка второго рода при проверке гипотезы, что изучаемый объект случай.

Относительный риск – отношение вероятности быть случаем при условии, что гипотеза "случай" принята, к вероятности быть случаем при условии, что эта гипотеза отвергнута $R = (a / (a + b)) : (e / (e + f))$.

Доля *ложноположительных*: $b / (a + b)$, т. е. доля не–случаев среди объектов, признанных случаями.

Доля *ложноотрицательных*: $e / (e + f)$, т. е. доля случаев среди объектов, признанных не–случаями.

Среди введенных характеристик только три независимых, остальные являются производными и могут быть получены из них простым пересчетом. Целесообразно выбирать в качестве основных частоту случаев, чувствительность и специфичность, или же частоту случаев и ошибки первого и второго рода.

2.4. Логические методы распознавания

Применение методов алгебры логики необходимо тогда, когда существенны не только количественные соотношения между величинами, характеризующими рассматриваемые процессы, но и связывающие их логические зависимости. При распознавании (диагностике) эти методы используют в случаях, когда отсутствуют сведения о количественном распределении объектов по

пространственным, временным или каким – то другим интервалам в соответствующем пространстве признаков, а имеются лишь детерминированные логические связи между анализируемыми объектами и их признаками.

Приведем примеры задач, для решения которых требуется применение методов алгебры логики: диагностика заболевания пациента на основе данных наблюдения и известных априорных зависимостей между видами заболеваний и соответствующими признаками; установление различных совокупностей признаков подлежащего диагностике заболевания, учет которых наряду с уже известными необходим для заключения о виде заболевания.

Для решения задач распознавания используется математический аппарат булевой алгебры. Он применяется для исчисления высказываний, установления зависимости и независимости высказываний, нахождения явного вида логической зависимости, а также для решения булевых алгебраических уравнений с одним (или более) неизвестных.

В логических системах распознавания классы и признаки объектов рассматриваются как логические переменные. Введем следующие обозначения для классов и признаков.

Пусть множество объектов подразделено на классы Ω_i, i = 1, …, m, а для описания объектов используются признаки $A_1, A_2, ..., A_n$.

Предположим, что вся априорная информация о классах объектов, выражающая: 1) связь между высказываниями $\Omega_1, ..., \Omega_m$ и $A_1, ..., A_n$; 2) зависимости между признаками $A_1, ..., A_n$; 3) зависимости между классами $\Omega_1, ..., \Omega_m$, представлена в форме булевых соотношений:

$$\begin{cases} \Omega_i = f_i(A_1, ..., A_n);\ \Omega_j = h_j(A_1, ..., A_n); \\ L_i(A_1, ..., A_n) \to \Omega_i; \\ F_i(A_1, ..., A_n; \Omega_1, ..., \Omega_m) = H_l(A_1, ..., A_n; \Omega_1, ..., \Omega_m); \\ \Phi_k(A_1, ..., A_n; \Omega_1, ..., \Omega_m) = 1; \\ y_r(A_1, ..., A_n) = 1; y_s(A_1, ..., A_n) = 1, ... \end{cases} \qquad (17)$$

Предположим, также, что наряду с (17) в результате эксперимента получены данные, касающиеся части признаков $A_1, ..., A_n$, характеризующих объекты

классов $\Omega_{1,...,}\Omega_{\boldsymbol{m}}$, и что эти данные выражены как булева функция $G(\boldsymbol{A}_1,...,\boldsymbol{A}_{\boldsymbol{n}}) =$ $= 1$.

Прямая задача распознавания [5] состоит в том, чтобы определить, какие выводы можно сделать относительно классов $\Omega_{1,...,}\Omega_{\boldsymbol{m}}$ на основе априорной информации (17) и апостериорной информации $G(\boldsymbol{A}_1,...,\boldsymbol{A}_{\boldsymbol{n}}) = 1$, т. е. требуется определить неизвестную функцию $\boldsymbol{F}(\Omega_1,...,\Omega_{\boldsymbol{m}})$, удовлетворяющую уравнению

$$\overline{\boldsymbol{G}}(\boldsymbol{A}_1,...,\boldsymbol{A}_{\boldsymbol{n}}) + \boldsymbol{F}(\Omega_1,...,\Omega_{\boldsymbol{m}}) = 1 \qquad (18)$$

при ограничениях (17).

Сопряженная задача заключается в том, чтобы установить, какие совокупности признаков $\boldsymbol{A}_1,...,\boldsymbol{A}_{\boldsymbol{n}}$ должны иметь место, если известны некоторые сведения о классах $\Omega_{1,...,}\Omega_{\boldsymbol{m}}$, т. е. требуется определить неизвестную функцию $G_1(\boldsymbol{A}_1,...,\boldsymbol{A}_{\boldsymbol{n}})$, удовлетворяющую уравнению

$$\overline{\boldsymbol{F}}_1(\Omega_{1,...,}\Omega_{\boldsymbol{m}}) + G_1(\boldsymbol{A}_1,...,\boldsymbol{A}_{\boldsymbol{n}}) = 1 \qquad (19)$$

при заданной функции $\boldsymbol{F}_1(\Omega_1,...,\Omega_{\boldsymbol{m}})$ и связях (18).

Обратная задача распознавания [5]заключается в том, чтобы определить множество априорно неизвестных посылок $G(\boldsymbol{A}_1,...,\boldsymbol{A}_{\boldsymbol{n}})$, из которых следуют некоторые выводы $\boldsymbol{F}(\Omega_1,...,\Omega_{\boldsymbol{m}})$ при условии, что признаки $\boldsymbol{A}_1,...,\boldsymbol{A}_{\boldsymbol{n}}$ и классы $\Omega_{1,...,}\Omega_{\boldsymbol{m}}$ связаны зависимостями (19).

Рассмотрим пример представления медицинских знаний с помощью операций математической логики. Допустим, что из медицинских наблюдений известны следующие связи: 1) признак A_1 появляется при диагнозе D_2; 2) если имеется диагноз D_1 и отсутствует диагноз D_2, то должен появляться признак A_2; 3) если появляются признаки A_1 или A_2 оба вместе, то может быть или диагноз D_1 или D_2 оба вместе.

Первое условие записывается в виде $D_2 \rightarrow A_1$, второе – $D_1 \wedge D_2 \rightarrow A_2$, третье – $A_1 \vee A_2 \rightarrow D_1 \vee D_2$.

Так как все эти условия справедливы одновременно, то они могут быть записаны в виде булевой функции событий:

$$E = [D_2 \rightarrow A_1] \wedge [D_1 \wedge D_2 \rightarrow A_2] \wedge [A_1 \vee A_2 \rightarrow D_1 \vee D_2] .$$

2. 5. Структурные методы распознавания

В настоящее время двумерные и многомерные изображения занимают видное место в качестве объектов распознавания (классификации). Например, техническое зрение роботов, медицинские системы обследования и диагностики – рентгенография, компьютерная томография, ангиография данные аэрофотосъемки и съемки со спутников, результаты неразрушающего контроля в промышленности и т. д.

Для решения задач распознавания изображений применяются структурные методы распознавания, называемые также лингвистическими и синтаксическими. Особенность этого метода заключается в том, что априорными описаниями классов являются *структурные описания* – формальные конструкции. Они используются при анализе иерархической структуры объекта и отношений, существующих между отдельными элементами этой структуры.

Во многих случаях апостериорная информация о распознаваемых объектах или явлениях содержится в записях соответствующих сигналов (электрокардиограмм, электроэнцефалограмм, инфракрасных сигналов, порождаемых исследуемыми органами, акустических сигналов функционирующих органов и т. д.). Традиционно для определения признаков (например, в задачах медицинской диагностики) используется разложение в ряды по ортогональным функциям. При этом в качестве признаков берутся коэффициенты разложения (в частности, ряда Фурье, полиномов Эрмита, Лежандпа, Чебышева, разложения Карунена–Лоэва и т. д.). Возможно использование в качестве признаков и некоторых характерных элементов экспериментальных кривых (точки минимума, максимума и др.), например, электрокардиограммы, содержащей типичные структурные элементы – зубцы (экстремальные точки) *P, Q, R, S, T,* связанные с циклами деятельности сердца.

Использование в качестве признаков характерных элементов экспериментальных кривых, их структуры базируется на том факте, что структура сигналов, отраженных распознаваемыми объектами или порожденных ими, однозначно определяется структурой наблюдаемого объекта.

В структурных методах распознавания описание объектов производится языковыми средствами. Правила языка, определяющие способы построения

объектов из непроизводных элементов, называют грамматикой. В соответствии с грамматикой объект представляется предложением на этом языке.

Процедура распознавания на основе использования структурных методов состоит из следующих этапов: предварительной обработки, описания или представления объекта и синтаксического анализа. На этапе предварительной обработки предъявленный для распознавания объект подвергается кодированию и фильтрации, восстановлению и улучшению качества.

Объект после предварительной обработки представляется некоторой структурой языкового типа (например, цепочкой или графом). Процесс получения представления объекта включает в себя процедуры: а) разбиения (сегментации) объекта; б) выделения признаков – непроизводных элементов. В результате каждый объект получает свое представление с помощью некоторого набора непроизводных элементов и ряда фиксированных синтаксических операций. Например, при использовании операции соединения объект получает представление в виде некоторой цепочки соединенных непроизводных элементов.

Система распознавания должна обнаруживать синтаксические связи, существующие в объекте. Решение о синтаксической правильности представления объекта, т. е. о принадлежности его к некоторому классу, задаваемому определенной синтаксической системой или грамматикой, вырабатывается *синтаксическим анализатором* (*блоком грамматического разбора*). При выполнении синтаксического анализа (грамматического разбора) анализатор воспроизводит полное синтаксическое описание объекта в виде дерева грамматического разбора, если соответствующий объект является синтаксически правильным. В противном случае объект либо отклоняется, либо подвергается анализу с помощью других заданных грамматик, которыми могут описываться другие классы изучаемых объектов.

Процедура распознавания – это сравнение с эталоном. Цепочка непроизводных элементов, представляющая анализируемый объект, сопоставляется с цепочками непроизводных элементов, описывающих классы. Распознаваемый объект с помощью выбранного *критерия согласия* (*подобия*) относится к тому классу, с которым обнаруживается наилучшая близость.

Реализация процесса распознавания на основе структурных методов

Для распознавания неизвестного объекта на основе структурных методов [5, 6] необходимо найти его непроизводные элементы и отношения между ними, а затем с помощью синтаксического анализа (грамматического разбора) установить, согласуется ли описание объекта с грамматикой, которая, по предположению, могла его породить.

Для формирования соответствующей грамматики можно воспользоваться либо априорными сведениями о распознаваемых объектах, либо результатами изучения конечного выборочного множества "типичных" объектов. В первом случае говорят о задании грамматики на основе эвристических соображений, во втором – о выводе грамматики.

Допустим, что мы имеем дело с двумя классами объектов (Ω_1 и Ω_2) и объекты, входящие в эти классы, можно описать с помощью признаков, принадлежащих некоторому конечному множеству. Эти признаки можно считать основными символами (элементы основного словаря) и обозначать через V (их называют также непроизводными символами, непроизводными элементами). Каждый объект может рассматриваться как цепочка или предложение, поскольку он составлен из элементов основного словаря.

Пусть также существует грамматика $\mathbf{\Gamma}$ такая, что порождаемый ею язык состоит из предложений (объектов), принадлежащих исключительно одному из классов, например, классу Ω_1. Эта грамматика может быть использована для классификации объектов, так как предъявленный неизвестный объект можно отнести к классу Ω_1, если он является предложением языка $L(\mathbf{\Gamma})$. В противном случае объект приписывается классу Ω_2. В данном случае объект попадает в класс Ω_2 исключительно потому, что он не входит в класс Ω_1: если оказывается, что объект не является грамматически правильным предложением в смысле грамматики $\mathbf{\Gamma}$, то предполагается, что он должен принадлежать классу Ω_2. На самом же деле объект может не относиться и к классу Ω_2. Поэтому необходимо располагать грамматиками $\mathbf{\Gamma_1}$ и $\mathbf{\Gamma_2}$, порождающими языки $L(\mathbf{\Gamma_1})$ и $L(\mathbf{\Gamma_2})$ соответственно. Объект попадает в тот класс, в языке которого он оказывается грамматически правильным предложением. Если последнее не выполняется, то очевидно, что данный объект не принадлежит ни одному из двух заданных классов и, следовательно, требуется еще одна грамматика $\mathbf{\Gamma_3}$ и т. д. В

случае m классов рассматривается m грамматик и связанных с ними языков $L(\mathbf{\Gamma}_i)$, $i = 1, \ldots, m$. Распознаваемый объект относится к классу Ω_i, если он является грамматически правильным предложением языка $L(\mathbf{\Gamma}_i)$. Если объект оказывается грамматически правильным предложением более чем одного языка, решение относительно его принадлежности принимается точно таким же образом, как это делается при использовании других методов распознавания, когда оказывается, что результаты реализации процедуры распознавания не позволяют определенно отнести объект к одному из заданных классов.

Практическое использование структурного метода требует обычно решения следующих основных проблем: 1) построение адекватного описания распознаваемых объектов; 2) выбора грамматики; 3) реализацию процесса распознавания с помощью процедур синтаксического анализа; 4) использования процедур обучения для вывода грамматик; 5) применения в рамках структурного подхода процедур из других методов распознавания, например, статистических для учета искажений случайного характера, кластерного анализа и т. д.

Основные приемы применения структурных методов распознавания рассмотрены в [1].

ЛИТЕРАТУРА

1. Журавлев Ю. И., Гуревич И. Б. Распознавание образов и распознавание изображений / Распознавание, классификация и прогноз. М.: Наука, 1989, вып. 2.

2. Айвазян С.А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. М.: Финансы и статистика, 1989.

3. Распознавание образов и медицинская диагностика / Под ред. Ю. Неймарка. М.: Наука, 1972.

4. Гублер Е. В. Вычислительные методы анализа и распознавания патологических процессов. Л.: Медицина, 1984.

5. Горелик А. Л., Скрипкин В. А. Методы распознавания. М.: Высшая школа, 1986.

Св. план 2001, поз. 24

Учебное издание

Авторы: Степанова Маргарита Дмитриевна,
Самодумкин Сергей Александрович,
Лемешева Татьяна Леонидовна

МАТЕМАТИЧЕСКИЕ МЕТОДЫ ДИАГНОСТИКИ В МЕДИЦИНСКИХ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМАХ

УЧЕБНО – МЕТОДИЧЕСКОЕ ПОСОБИЕ

по курсу "Прикладные интеллектуальные системы и системы принятия решений"
для студентов специальности Т 10.04.00 "Искусственный интеллект"

Редактор Т. Н. Крюкова

Подписано в печать. . .04.04.2001. Формат 60x84 1/16.

Бумага . .офсетная. Печать ризографическая. Усл. печ л. 2,79.

Уч.-изд. л. 2,5 Тираж 180 экз. Заказ 223.

Учреждение образования “Белорусский государственный университет информатики и радиоэлектроники”

Отпечатано в БГУИР. Лицензия ЛП № 156. 220013, Минск, П. Бровки, 6