ОБЪЕДИНЕННЫЙ ИНСТИТУТ ЯДЕРНЫХ ИССЛЕДОВАНИЙ

ЛАБОРАТОРИЯ ВЫЧИСЛИТЕЛЬНОЙ ТЕХНИКИ И АВТОМАТИЗАЦИИ

На правах рукописи

БУЛЫГИН ВАЛЕНТИН ПЕТРОВИЧ

НЕКОТОРЫЕ ЗАДАЧИ СТАТИСТИЧЕСКОЙ КЛАССИФИКАЦИИ МНОГОМЕРНЫХ НАБЛЮДЕНИЙ В УСЛОВИЯХ ОГРАНИЧЕННОГО ОБЪЕМА ОБУЧАЮЩИХ ВЫБОРОК

Специальность 01.01.10 – математическое обеспечение вычислительных машин и систем

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

ДУБНА 1978 г.

ОБЪЕДИНЕННЫЙ ИНСТИТУТ ЯДЕРНЫХ ИССЛЕДОВАНИЙ

ЛАБОРАТОРИЯ ВЫЧИСЛИТЕЛЬНОЙ ТЕХНИКИ И АВТОМАТИЗАЦИИ

На правах рукописи

БУЛЫГИН ВАЛЕНТИН ПЕТРОВИЧ

НЕКОТОРЫЕ ЗАДАЧИ СТАТИСТИЧЕСКОЙ КЛАССИФИКАЦИИ МНОГОМЕРНЫХ НАБЛЮДЕНИЙ В УСЛОВИЯХ ОГРАНИЧЕННОГО ОБЪЕМА ОБУЧАЮЩИХ ВЫБОРОК

Специальность 01.01.10 – математическое обеспечение вычислительных машин и систем

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата физико-математических наук

ДУБНА 1978 г.

Работа выполнена во ВНИИ медицинского приборостроения Министерства медицинской промышленности

НАУЧНЫЙ РУКОВОДИТЕЛЬ: доктор технических наук М.А.Шнепс-Шнеппе

ОФИЦИАЛЬНЫЕ ОППОНЕНТЫ: доктор физико-математических наук Питьев Ю.П.

кандидат физико-математических наук Деев А.Д.

ВЕДУЩЕЕ ПРЕДПРИЯТИЕ: ВЦ АН СССР

Защита состоится " "_____" 1978г. на заседании специализированного совета Д047.01.04 при Лаборатории вычислительной техники и автоматизации ОИЯИ.

Автореферат разослан " "_____" 1978г..

С диссертацией можно познакомиться в библиотеке ОИЯИ.

Ученый секретарь специализированного совета, к.ф.-м.н. /Пузынина Т.П./

Пузынина

Актуальность работы. Диссертация выполнена в рамках исследований по разработке математического проблемно-ориентированного обеспечения для расчета автоматизированных медицинских приборов и устройств для массовых обследований населения. Схема расчета и построения автоматизированных приборов выглядит следующим образом (Смирнов И.П., Шнепс-Шнеппе М.А., 1972 г.)

- на основании исходного перечня подлежащих выявлению заболеваний для различных половозрастных и профессиональных групп населения ведущими медицинскими учреждениями собирается исходный статистический материал;

- собранные по полной медицинской программе обследования истории болезни обрабатываются на мощных ЭВМ. Основная цель обработки - получение решающих правил и отсеивание неинформативных (для заданной классификации по группам риска) признаков;

- решающие правила реализуются затем в вычислительном блоке автоматизированного устройства (микро - ЭВМ, микропроцессоре);

- с помощью решающих правил для каждого нового обследуемого пациента прогнозируется вероятность наличия у него данного заболевания или группы родственных заболеваний. Лица с неблагоприятным прогнозом объединяются в группы риска и направляются на соответствующее детальное обследование.

Описанный подход по автоматизации переработки многомерной информации в медицинских приборах и устройствах, осуществим при выполнении ряда условий, важное место среди которых занимает создание математического обеспечения больших ЭВМ, ориентированного на решение задач статистической классификации многомерных наблюдений. Среди зарубежных аналогов такого обеспечения можно назвать систему программ OLPARS, пакеты программ BMDP, систему программ PROMENADE, пакет программ SPSS и др..

Цель работы. Основной целью автора диссертации было получение теоретических результатов, направленных на усиление алгоритмических возможностей создаваемого во ВНИИ медицинского приборостроения аналогичного математического обеспечения ЕС ЭВМ (разрабатываемого по заданию I0.07 проблемы 0.80.I4 координационного плана ГКНТ), в том числе решение вопросов выбора информативного набора показателей при классификации многомерных наблюдений в одну из нескольких совокупностей в условиях существенно ограниченного объема обучающего статистического материала. Для этого было необходимо:

а) определить пути прохождения матрицы данных в проблемно-ориентированном пакете программ для решения задач статистической классификации;

б) разработать метод описания совместного расположения классифицируемых групп данных, коррелированный с ошибками классификации;

в) разработать методы оценки качества классификатора, работающего в условиях числа классифицируемых групп больше двух и существенно ограниченного объема обучающих выборок;

г) на этой основе разработать правила остановки различных пошаговых алгоритмов (step-wise) выделения информативного набора показателей;

д) разработать алгоритмическое ядро пакета программ и проверить с его помощью правильность закладываемых решений, оценить набор программных модулей для решения медико-технических задач.

Научная новизна. а) По сравнению с классическими схемами построения информативного набора (реализованными, например, в BMDP и SPSS) в диссертации изложен метод, который использует правила остановки, зависящие от объемов выборок и размерности пространства, что существенно расширяет класс обрабатываемых данных.

б) В работе введена характеристика взаимного расположения групп данных - матрица информативностей V^2, найдено распределение выборочной оценки этой матрицы, получены асимптотические разложения математического ожидания матрицы информативностей на выборочные значения дискриминантных функций, построены критерии информативности и правила остановки различных (step-wise) алгоритмов выделения информативного набора показателей.

в) Разработано алгоритмическое ядро пакета программ, ориентированного на решение задач статистической классификации.

Практическая ценность. Полученные результаты могут быть использованы в различных задачах классификации многомерных наблюдений, в частности, для расчета автоматизированных приборов и устройств, осуществляющих сложную интерпретацию медико-биологической информации (автоматический анализатор стадий наркоза по параметрам ЭЭГ, автоматизированные устройства для выявления групп риска больных различными заболеваниями и др.)

Реализация. Развитые методы построения информативного набора реализованы в системах программ для ЭВМ второго поколения: "Диагностика-2" (для ЭВМ БЭСМ-4) и "Диагностика-2бис" (для ЭВМ МИНСК-22), сданных в Государственный фонд алгоритмов и программ, а также являются составной частью работ по созданию пакета программ для ЕС ЭВМ, ориентированного на решение задач классификации многомерных наблюдений.

С помощью разработанного алгоритмического ядра пакета программ - систем программ "Диагностика-2" произведен расчет автоматизированного устройства для выявления групп риска больных ревматическими заболеваниями при массовых обследованиях населения.

При активном участии автора разработано "Обоснование разработки автоматизированного устройства для скрининга больных ревматическими заболеваниями". В настоящее время на основе отечественной микро-ЭВМ создан экспериментальный макет этого автоматизированного устройства. С помощью макета устройства обследовано около 1500 человек. Результаты обследования показали высокую эффективность заложенных в него решающих правил (82% правильных решений).

Апробация работы. Результаты работы докладывались на семинарах ЦЭМИ АН СССР и МВТУ им.Баумана, ЛВТА ОИЯИ, а также на Всесоюзных конференциях: по математическому обеспечению ЭВМ в медико-биологических исследованиях в г.Обнинске (1976) и по применению многомерного статистического анализа и в экономике и оценке качества продукции в г.Тарту (1977).

Объем работы. Диссертация представлена на 120 страницах и состоит из введения, 4 глав, заключения и приложения.

Содержание работы. Создание математического обеспечения ЭВМ для решения задач статистической классификации в режиме off-line имеет два основных аспекта: 1) системные вопросы хранения и выборки данных, программного управления заданиями с передачей управления из одного модуля в другой и др. 2) алгоритмический аспект, связанный с построением эффективных процедур классификации, выбора информативного набора и т.п. И если при создании проблемно-ориентированного пакета простая адаптация управляющих программ пакетов типа BMDP и SPSS не вызывает особых возражений, то алгоритмический состав модулей не соответствует современным статистическим представлениям анализа данных (analysis data). Мы имеем в виду учет "эффектов больших размерностей" при выборе информативного подмножества признаков и построения классификатора.

Не вдаваясь в подробную систематизацию алгоритмов, осуще-

ствляющих отнесение вновь поступившего измерения к одному из классов, условно, можно выделить два основных направления статистической классификации: непараметрические (локальные) методы, которые не связаны с видом плотностей распределения признаков, и параметрические, в которых вид плотности известен с точностью до параметров. На практике наиболее часто используемой моделью является многомерная нормальная плотность или смесь нормальных плотностей, а проблемы построения классификатора сконцентрированы в основном вокруг линейных дискриминантных функций (Н.Г.Загоруйко, 1972). Такой способ поиска классификатора дает возможность постановки и решения ряда задач статистической классификации: прогнозирование качества классификации на "экзамене" (Lachenbruch P, 1968 г., Деев А.Л., 1972, Блохов Ю.С., 1974 г.), асимптотические разложения распределения ряда статистик дискриминантного анализа в условиях ограниченного объема обучающих выборок (M.Okamoto, 1938 г., Деев А.Л., 1970 г., Мешалкин Л.Д., 1971 г, Архаров Л.И., 1972 г.), оценка зашумленности обучающих выборок (Раудис Ш., 1975). Настоящая работа выполнена в рамках параметрического подхода.

Рассматривается задача классификации p-мерного наблюдения в одну из m нормальных совокупностей $H_i \sim N_p(\mu_i, \Sigma), i=\overline{1,m}$ с общей матрицей ковариаций. Параметры совокупностей предполагаются неизвестными и информация о них задается выборками $\{X_\alpha^{(i)}\}$, $\alpha=\overline{1,n_i}$, $i=\overline{1,m}$.

Выражение для вероятностей ошибочной классификации при использовании дискриминантной функции Фишера Р. (Андерсон Т., 1963 г.) содержит неизвестные параметры $\mu_i (i=\overline{1,m})$ и Σ. Замена этих параметров на их несмещенные оценки (т.е. $\hat{\mu}_i$ и $\hat{\Sigma}$) приводит к значительному смещению оценки ошибки и создает иллюзию высокого качества дискриминации. Для случая двух классифицируемых совокуп-

ностей /m=2/ получены различные асимптотические разложения математического ожидания ошибок классификации в терминах объемов выборок, размерности пространства признаков и расстояния Махаланобиса T^2 между генеральными совокупностями. Одним из главных достоинств расстояния T^2 для дискриминантного анализа является монотонная зависимость ошибок классификации от расстояния Махаланобиса (в случае равных априорных вероятностей). Такая зависимость оправдывает использование T^2 как меры информативности набора показателей.

Для дискриминации m совокупностей существует несколько различных мер информативности (мер расстояний) (С.Рао, 1949 г., С.Уилкс, , 1969 г.). Однако связь цитированных критериев Рао и Уилкса с ошибками классификации в форме аналогичной соотношению для двухклассового случая не прослеживается, что затрудняет прогнозирование ошибок классификации и построение информативного подмножества признаков. В работе предложена и изучается статистика, являющаяся естественным обобщением статистики Хотеллинга и более просто связанной с ошибками классификации.

<u>В первой главе</u> кратко анализируется путь прохождения матрицы данных через пакет программ для решения задач статистической классификации. Важным аспектом анализа матрицы данных является редукция информации с учетом размеров матрицы данных и числа классифицируемых групп данных. Во § 2 первой главы формализуется описание групп данных и вводится основной инструмент исследования, естественное обобщение расстояния Махаланобиса T^2 – матрица информативностей V^2 и изучаются ее основные свойства.

Пусть в евклидовом пространстве $\mathcal{X}_p$ размерности p заданы m нормальных совокупностей H_i $(i=\overline{1,m})$ с векторами средних μ_i $(i=\overline{1,m})$ и одинаковой для всех совокупностей матрицей ковариаций Σ. Введем неотрицательно-определенную матрицу информативностей V^2 разме-

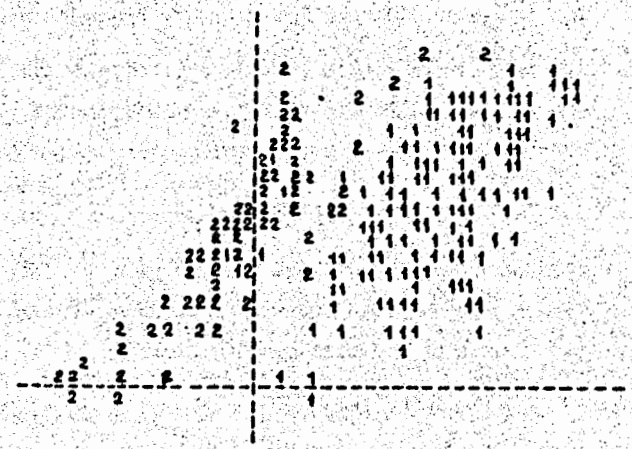

Рис. I. Отображение 2 групп данных (по ревматическим заболеваниям)

ра $(m-1)\times(m-1)$ и будем рассматривать только матрицы ранга $m-1$:

$$V^2_{ij} = (\mu_i-\mu_m)'\Sigma^{-1}(\mu_j-\mu_m), \qquad i,j=\overline{1,m-1},\ p\geq m-1. \tag{I.I}$$

Матрица V^2 инвариантна относительно невырожденного линейного преобразования в $\mathcal{X}_p$, а определитель $\det V^2$ является квадратом $(m-1)$ – объема (в метрике, задаваемой матрицей Σ).

Вероятность $P(i/m)$ отнести объект из совокупности H_i к совокупности H_m равна интегралу по отрицательному квадранту от $m-1$-мерной нормальной плотности:

$$P(i/m)=\int_{-\infty}^{0}\cdots\int dz_1 dz_2 \ldots dz_{m-1}\, N_{m-1}(\vartheta_i, A^2). \tag{I.2}$$

Вектор средних ϑ_i и матрица A^2 в подинтегральном выражении определяется через элементы матрицы V^2:

$$(\hat{V}_i)_k = V_{ik}^2 - \frac{1}{2}V_{kk}^2, \quad k = \overline{1,m-1}; \qquad A^2 = V^2. \tag{I.3}$$

Матрица информативностей просто связана с критериями Рао R и Уилкса-Λ. Пусть элементы положительно-определенной матрицы $\mathcal{D}'\mathcal{D}$ порядка $(m-1)\times(m-1)$, ранга $m-1$ определяются соотношениями:

$$(\mathcal{D}'\mathcal{D})_{ij} = m\delta_{ij} - 1; \quad i,j = \overline{1,m-1}, \tag{I.4}$$

где δ_{ij} - дельта функция Кронекера. Тогда след матрицы $\mathcal{D}'V^2\mathcal{D}$ суть критерий Рао, т.е. сумма всех взаимных расстояний Махаланобиса, а для критерия Уилкса справедливо соотношение:

$$\Lambda = 1 / \det\left(J + \frac{1}{m-1}\mathcal{D}'V^2\mathcal{D}\right); \tag{I.5}$$

Рассмотрим выборочные свойства матрицы V^2. Пусть из m p-мерных нормальных совокупностей с векторами средних μ_i и одинаковой для всех совокупностей матрицей ковариаций Σ извлечены выборки объема n_i $(i=\overline{1,m})$, соответственно. Пусть $\hat{\mu}_i$ и $\hat{\Sigma}$ оценки параметров совокупностей. Тогда распределение элементов выборочной оценки матрицы V^2 дается теоремой:

Теорема I.

Пусть элементы положительно-определенной матрицы $C'C$ порядка $(m-1)\times(m-1)$ ранга $m-1$ зависят только от объемов выборок и определяются соотношениями:

$$(C'C)_{ij} = n_i\delta_{ij} - \frac{n_i n_j}{k+m}; \quad k = \sum_{i=1}^{m} n_i - m. \tag{I.6}$$

Тогда случайная матрица $\frac{1}{k}C'\hat{V}^2C$ может быть предоставлена в виде

$$\frac{1}{k}C'\hat{V}^2C = \mathcal{U}_1^{\frac{1}{2}}\mathcal{U}_2^{-1}\mathcal{U}_1^{\frac{1}{2}}, \tag{I.7}$$

где случайные матрицы $\mathcal{U}_1$ и $\mathcal{U}_2$ взаимно независимы и распределены

$$\mathcal{U}_1 \sim W_{m-1}(p, J, C'V^2C),$$

как нецентральное $(m-1)$-мерное распределение Уишарта с числом степеней свободы p и "параметром нецентральности" матрицей $C'V^2C$;

$$\mathcal{U}_2 \sim W_{m-1}(k-p+m-1, J),$$

как центральное распределение Уишарта с числом степеней свободы $k-p+m-1$, J -единичная матрица.

Замечание I. Распределение матрицы $\frac{1}{k}C'\hat{V}^2C$ является многомерным аналогом f-распределения Хотеллинга, в духе распределений изучавшихся Г.П.Климовым (1973 г.). Для случая $V^2 \neq 0$ (проверка гипотезы о том, что все выборки извлечены из одной совокупности), плотность распределения выписывается в явной форме. Если $m=2$, то распределение переходит в известное распределение Хотеллинга (С.Рао, 1968 г.):

$$\frac{c\hat{T}^2}{k} \sim \frac{\chi^2(p, cT^2)}{\chi^2(k-p+1)},$$

где $c = \frac{n_1 n_2}{n_1+n_2}$, $k = n_1+n_2-2$.

Замечание 2. Математическое ожидание матрицы $\hat{V}^2$ является смещенной оценкой матрицы V^2:

$$E(\hat{V}^2) = \frac{k}{k-p+1}\left(V^2 + p(C'C)^{-1}\right). \tag{I.8}$$

Вторая глава посвящена получению асимптотических разложений математического ожидания некоторых многоклассовых статистик, связанных с матрицей информативностей V^2. Эти разложения дают возможность оценить качество классификатора в условиях ограниченного объема обучающих выборок.

Для двух нормальных классифицируемых совокупностей М.Окамото (1963 г.) получил асимптотическое разложение математического ожидания вероятностей ошибочной классификации как функция p, n_1, n_2

и расстояния Махаланобиса с точностью до членов $\frac{1}{t^2}$, где t одно из чисел n_1, n_2, n_1+n_2-2. Анализ остаточного члена в разложении Окамото и ряд особенностей поведения некоторых статистик дискриминантного анализа (Ю.Н.Благовещенский, I972) показывают, что обычные асимптотические свойства статистических процедур нарушаются, когда размерность пространства и объемы выборок сравнимы $\frac{p}{n_i} \to \lambda_i = const$, $p \to \infty$, $n_i \to \infty$, $i=1,2$. В задаче классификации двух нормальных совокупностей А.Д.Деевым (I972) получено асимптотическое разложение математического ожидания ошибок в этой асимптотике. Иной подход был предложен в работе И.С.Енюкова (I974), где методом прямого вероятностного рассуждения оценивалось математическое ожидание расстояния Махаланобиса между проекциями совокупностей на выборочные значения дискриминантных векторов в той же асимптотике. В задаче классификации больше двух нормальных совокупностей ($m>2$), следуя цитированной работе И.С.Енюкова, будем оценивать математическое ожидание матрицы $V^2(\hat{w})$ проекций совокупностей на выборочные значения дискриминантных векторов с последующей подстановкой полученных выражений в формулы (I.2) и (I.3). Дискриминантные направления в исходном пространстве признаков задаются набором p-мерных векторов:

$$\hat{w}_i = (\hat{\mu}_i - \hat{\mu}_m)'\hat{\Sigma}^{-1}; \quad i = \overline{1, m-1} \qquad (2.1)$$

Построим из этих векторов матрицу $\hat{W}$ размерности $p \times (m-1)$, столбцы которой соответствуют векторам $\hat{w}_i$. Аналогично, из p-мерных векторов $(\mu_i - \mu_m)$, $i=\overline{1,m-1}$ построим матрицу M размерности $p \times (m-1)$. Тогда элементы матрицы $V^2(\hat{w})$ в дискриминационном пространстве задаются соотношениями:

$$V^2(\hat{w}) = (M'\hat{W})(\hat{W}'\Sigma\hat{W})^{-1}(\hat{W}'M),$$

Отметим важное свойство матрицы $V^2(\hat{w})$: если ранг матриц $\hat{W}$ и M равен $p = m-1$, то $V^2(\hat{w}) = V^2$, что соответствует проектированию совокупностей самих в себя.

Теорема 2.

Пусть ранг матрицы V^2, построенной для истинных значений параметров нормальных совокупностей, равен $m-1$. Пусть объемы выборок $n_i(N) \to \infty$, $N \to \infty$, $i = \overline{1,m}$ так, что $\frac{n_i(N)}{n_j(N)} \to \lambda < \infty$ для всех i и j, $i,j = \overline{1,m}$. Тогда для всех положительно-определенных матриц V^2 таких, что справедливо матричное неравенство $V^2 \geq J$ и размерности $p \geq m-1$ математическое ожидание матрицы $V^2(\hat{w})$ дается выражением

$$E(V^2(\hat{w})) = G^2 + O(\tfrac{1}{t}) = V^2\left(1 - \frac{p-m+1}{k}\right) - (p-m+1)(C'C)^{-1} + O(\tfrac{1}{t}), \qquad (2.2)$$

где t одно из чисел n_1, n_2, ..., n_m, K, а K - число степеней свободы распределения $\hat{\Sigma}$, равное $\sum_{i=1}^{m} n_i - m$.

Замечание. Разложение (2.2) является аналогом разложения Окамото, но построенное для меры взаимного расположения совокупностей, а не для ошибок.

Анализ остаточного члена разложения (2.2) показывает, что при больших размерностях p, его величина оказывается сравнимой с главным членом разложения (размерность входит в качестве множителя в остаточный член).

Теорема 3.

Пусть проектирование в дискриминационное пространство осуществляется набором дискриминантных векторов

$$\hat{w}_i = (\hat{\mu}_i - \hat{\mu}_m)'\Sigma^{-1}, \quad i = \overline{1, m-1}$$

(т.е. матрица ковариаций предполагается известной).

Пусть целые p, n_1, n_2, ..., $n_m \to \infty$, так что элементы положительно-определенной матрицы $p(C'C)^{-1}$ остаются ограниченными и

и $p \geq 2(m-1)$. Тогда для всех положительно-определенных матриц V^2 ранга $m-1$ справедливо

$$E(V^2(\hat{w})) = G^2 + O(\tfrac{1}{t}) = V^2(V^2 + p(c'c)^{-1})^{-1}V^2 + O(\tfrac{1}{t}); \quad (2.3)$$

где t одно из чисел $n_1, n_2, \ldots, n_m$.

Замечание I. Выражение для главного члена разложения получено в предположении, что матрица ковариаций известна. Для практических целей поправку на оценку матрицы $\hat{\Sigma}$ можно извлечь из разложения (2.2) и использовать выражение вида

$$G^2 = \frac{k-p}{k-m+1} V^2(V^2 + p(c'c)^{-1})^{-1}(V^2 + (m-1)(c'c)^{-1}). \quad (2.4)$$

Замечание 2. В формулах (2.2) и (2.3) участвует неизвестный параметр - матрица V^2. Согласно рекомендациям П.Лахенбруха (I968 г.) по технике OS, на практике следует подставить несмещенную оценку этой матрицы:

$$V^2_{несм} = \frac{k-p+1}{k} V^2 - p(c'c)^{-1}. \quad (2.5)$$

Однако, существует опасность получения отрицательно-определенной матрицы либо в качестве несмещенной оценки (2.5), либо для матрицы главного члена разложений (2.2) и (2.3). В работе П.Лахенбруха и Мики (I969 г.) в аналогичной ситуации для разложения Окамото предлагается заменять значение расстояния Махаланобиса на оценку вида $\frac{k-p+1}{k}\hat{T}^2$. Для разложений (2.2) и (2.3) можно рекомендовать в этом случае заменять матрицу V^2 на оценку $\frac{k-p+1}{k}\hat{V}^2$.

В третьей главе на основе полученных результатов строится критерий информативности показателей и правило остановки алгоритмов выделения информативного набора.

Решающим критерием информативности набора показателей является, разумеется, суммарная ошибка классификации. Трудности вычислительного характера заставляют обратиться к обобщенным мерам расстояний между совокупностями. Зависимость (I.2) ошибок классификации и знание распределения (I.7) дают определенное основание использовать инварианты матрицы $\frac{1}{k}c'\hat{V}^2c$ в качестве мер информативности отдельного показателя. В данной работе рассматривались след и определитель этой матрицы.

Утверждение I. Пусть $\hat{R}$ взвешенная сумма всех взаимных расстояний Махаланобиса - выборочный аналог критерия Рао (I949):

$$\hat{R} = \frac{1}{k}\sum_{i\neq j=1}^{m} \frac{n_i n_j}{k+m}\hat{T}^2_{ij}.$$

Пусть $(c'\hat{V}^2c)_{ii} = i$ - диагональный элемент матрицы $c'\hat{V}^2c$. Тогда статистика $\hat{R}$ распределена как сумма независимых случайных величин имеющих нецентральное f -распределение Хотеллинга:

$$\hat{R} = \frac{1}{k}Sp\,c'\hat{V}^2c \sim \sum_{i=1}^{m} f_i(p, k-p+1, (c'V^2c)_{ii}). \quad (3.I)$$

Утверждение 2. Определитель матрицы $\frac{1}{k}c'\hat{V}^2c$ распределен как отношение двух независимых случайных величин $\mathcal{D} = \frac{t_1}{t_2}$ таких, что

$$t_2 \sim \prod_{i=1}^{m-1} f^2(k-p-i+2)$$

а t_1, распределена как определитель матрицы, совместным распределением элементов которой является нецентральное распределение Уишарта $W_{m-1}(p, J, c'V^2c)$. Это позволяет, в частности, выявить характер смещенности $\det\hat{V}^2$ как оценки $\det V^2$.

Разность значений $\hat{R}$ и $\hat{\mathcal{D}}$ на размерностях пространства p и $p-1$ может быть использована для оценки информативности отдельного показателя.

При использовании пошаговых (step-wise) процедур построения информативного набора показателей, чрезвычайно полезной оказывается связь между критериями $\hat{R}$ и $\hat{\mathcal{D}}$, с одной стороны, и коэффици-

ентами дискриминантных функций, с другой. Эта связь позволяет существенно сократить время анализа информативности показателей и дается следующей

Леммой. Пусть $\hat{R}$ и $\hat{D}$ значения критериев, построенных по размерности пространства p, пусть $\hat{R}(-s)$ и $\hat{D}(-s)$ значения соответствующих критериев построенных по размерности $p-1$, с исключением признака номер s. Пусть $u(s)$ вектор столбец размерности $m-1$, построенный из s-ых компонент дискриминантных функций. Тогда изменение критериев $\hat{R}$ и $\hat{D}$ при исключении признака номер s даются формулами

$$\Delta R(s) = \hat{R} - \hat{R}(-s) = \frac{1}{k}\frac{1}{\sigma^{ss}} u'(s) C'C\, u(s);$$

$$\Delta D(s) = \hat{D} - \hat{D}(-s) = \frac{\det C'\hat{V}^2 C}{k^{m-1}\sigma^{ss}} u'(s)(C'\hat{V}^2 C)^{-1} u(s); \qquad (3.2)$$

здесь σ^{ss} - s-ый диагональный элемент матрицы $\hat{\Sigma}^{-1}$.

Важным аспектом работы различных алгоритмов построения информативногонабора, таких как последовательного удаления (step-down), последовательного присоединения (step-up) и др. является правило остановки. Такое правило, естественно, связывать с качеством классификации при использовании данного набора признаков.

Утверждение 3. В предположениях теоремы 3 математическое ожидание следа матрицы $C'V^2(\hat{\omega})C$ дается выражением

$$E[Sp(C'V^2(\hat{\omega})C)] = SpG^2 + O(\tfrac{1}{t}) =$$

$$= Sp[C'V^2C(C'V^2C + pJ)^{-1}C'V^2C] + O(\tfrac{1}{t}), \qquad (3.3)$$

где t одно из чисел $n_1, n_2, \ldots, n_m$.

Для любых p, $SpG^2 \le SpC'V^2C$ и кроме того у величины SpG^2 возможно существование максимума (что можно трактовать как недостаточность информации в обучающих выборках). Достижение этого максимума может служить правилом остановки различных алгоритмов выделения информативного набора.

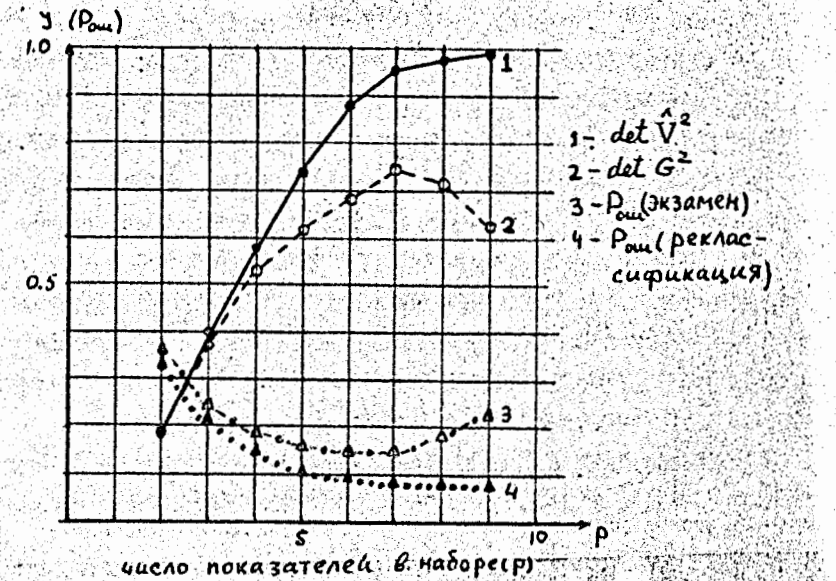

Рис. 2. Поведение критериев информативности набора в задаче автоматического анализа глубины наркоза по параметрам ЭЭГ.

Величину $\det G^2$ также удобно рассматривать как меру дискриминантной способности набора p показателей (хотя, быть может, более корректно в качестве критерия использовать математическое ожидание величины $\det(C'V^2(\hat{\omega})C))$. Определитель матрицы $\det G^2$ обладает свойствами аналогичными свойствам следа и также может быть использован для остановки алгоритмов построения информативного набора.

В четвертой главе дана краткая характеристика системы программ "Диагностика-2" (для ЭВМ БЭСМ-4), которая является алгоритмическим ядром пакета программ для ЕС ЭВМ и приведены примеры расчета с помощью этой системы автоматизированного устройства для выявления групп риска больных ревматическими заболеваниями и рас-

чета автоматического анализатора глубины наркоза.

Система программ "Д-2" представляет набор программ, объединенных одной задачей: построение решающего правила (классификатора) и нахождение информативного набора признаков в задачах статистической классификации. Максимальные размеры матрицы данных по одной задаче определяются условиями: $p \leq 125$, $\sum n_i \leq 2$[illegible] векторов (истории болезни). Максимальное число классифицируемых групп данных - I5 ($m = 15$). В системе "Д-2" реализован алгоритм последовательного удаления (step-down). В качестве критерия информативности отдельного признака используется определитель $\det \hat{V}^2$. Правило остановки основано на достижении максимума $\det G^2$.

С помощью системы "Д-2" можно совершать следующие действия: присваивать "имя" шифр вводимому статистическому материалу и хранить его на внешних носителях информации; получать статистические параметры распределений с коррекцией на неизмеренность ряда показателей; строить линейные решающие правила; проводить контроль качества классификации; упорядочивать набор признаков по информативности и изменять размерность пространства ("по умолчанию" исключается один наименее информативный по критерию $\det V_i^2$, либо по желанию пользователя исключается группа признаков), находить информативный набор; осуществлять вывод информации в виде таблиц, графиков, картинок выборочных гроздей, комментариев и т.п..

С помощью данного математического обеспечения проводился анализ целого ряда наборов медицинских данных и, в частности, расчет автоматизированного диагностического устройства для выявления ревматических заболеваний при массовых обследованиях населения. Данные для анализа были собраны под руководством Института ревматизма АМН СССР, при участии Рижского медицинского института.

та. Размеры матрицы данных: общий объем выборки $\sum n_i = 950$, причем объемы выборок по классам варьировались от 99 до 240, размерность пространства признаков $p = 123$, число классифицируемых групп данных $m = 5$. Первый и второй класс представляли группу "повышенного риска" (активные процессы) больных ревматическими заболеваниями, третий и четвертый класс - группу "риска" (неактивные процессы), и пятый класс - группу "здоровых". Суммарная ошибка реклассификации на исходной размерности пространства составила 9%. Изменение размерности пространства производилось по алгоритму step-down. Построение информативного набора по критерию $\det V^2$ показало существование явного максимума у величины $\det G^2$. Суммарная ошибка реклассификации на размерности $p = 60$ составила I6%. Суммарная ошибка на "экзамене" (I500 н/б) составила 22%.

Выводы

I. Важным аспектом анализа матрицы данных в проблемно-ориентированном пакете программ для решения задач статистической классификации является редукция информации с учетом размеров матрицы данных и числа классифицируемых групп данных. Для решения такого рода задач весьма продуктивными оказываются параметрические методы статистической классификации, основанные на многомерных нормальных распределениях.

2. Введена характеристика взаимного расположения групп данных, естественное обобщение расстояния Махаланобиса - матрица информативностей V^2, коррелированная с ошибками классификации. Получено распределение выборочной оценки этой матрицы - многомерный аналог f-распределения Хоттелинга.

3. Качество классификатора, работающего в условиях ограниченного объема обучающих выборок можно оценивать по элементам матрицы

G^2 - математическому ожиданию матрицы информативностей, построенной для проекций генеральных совокупностей на выборочные значения дискриминантных функций. Получены асимптотические разложения этой матрицы в асимптотике больших объемов выборок (асимптотике Окамото) и в асимптотике больших размерностей (асимптотике Колмогорова-Деева).

4. На основе инвариантов матрицы информативностей V^2 и матрицы G^2 строятся критерии информативности показателей и правила остановки пошаговых процедур и построения информативного набора. Правила остановки зависят не только от взаимного расположения групп данных, но и от размеров матрицы данных, количества классифицируемых групп, числа элементов в каждой группе.

5. Развитые методы анализа матрицы данных реализованы в системе программ "Диагностика-2" для ЭВМ второго поколения (БЭСМ-4). Практический опыт работы с системой "Д-2" позволил рассматривать ее как алгоритмическое ядро пакета программ, ориентированного на решение задач статистической классификации. На системе "Д-2" произведен расчет автоматизированного устройства для выявления групп риска больных ревматическими заболеваниями. С помощью экспериментального макета этого устройства обследовано около 1500 человек. Результаты обследования показали высокую эффективность заложенных в него решающих правил (82% правильных решений).

Основные результаты диссертации опубликованы в следующих работах автора:

1. "Некоторые вопросы практического применения дискриминантного анализа". "Новости медицинской техники", М., вып.3, 1975 (в соавторстве с И.С.Енюковым).

2. "Построение ЛДФ в случае произвольного числа классифицируемых совокупностей". "Новости медицинской техники", М., вып.3, 1975.

3. Методы построения "экономичных" дискриминантных "функций". В сб. "Материалы по математическому обеспечению и использованию ЭВМ в медико-биологических исследованиях", Обнинск. 1976 (в соавторстве с Коноваловой Л.А.)

4. Поведение ошибок классификации при разделении многих классов в условиях ограниченного объема выборок". В сб. "Исследования по вероятностно-статистическому моделированию реальных систем". М., ЦЭМИ АН СССР, 1977.

5. Выбор информативного набора признаков в случае числа классов большего двух и ограниченного объема выборок". В сб. "Исследования по вероятностно-статистическому моделированию и реальных систем". М., ЦЭМИ АН СССР, 1977.

6. Выделение информативного подмножества признаков в многоклассовой задаче дискриминантного анализа. "Тезисы докладов Всес. научно-технической конф. "Применение многомерного статистического анализа в экономике и оценке качества продукции", г.Тарту, 1977.

7. Система программ "Диагностика-2". Алгоритмы и программы. Информ.бюллетень. М. ВНТИЦЕНТР, 1978, № 3. (в соавторстве с Лейбовским М.А. и Кацевой В.П.).

ЗАК 290 Л-94068 ТИР 120 ОТ 8.08.78
ВНИИМП