МАШИННЫЕ МЕТОДЫ ОБНАРУЖЕНИЯ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
1981 год Выпуск 88

УДК 519.95:681.3.06

ИНФОРМАЦИОННЫЙ АНАЛИЗ МАШИННОГО ОБНАРУЖЕНИЯ ЗАКОНОМЕРНОСТЕЙ (НА ПРИМЕРЕ ЗАДАЧИ РАСПОЗНАВАНИЯ ОБРАЗОВ)

Ш.Ю. Раудис

Введение

Одной из основных задач при анализе и синтезе методов машинного обнаружения закономерностей является установление условий, при которых эмпирические закономерности могут быть найдены [1]. При обработке таблиц экспериментальных данных типа "объект-свойство" обычно требуется установить связь между выходной переменной y и вектором входных параметров X - $y = f(X)$. При установлении конкретной эмпирической закономерности $y = f_\alpha(X,A)$ решаются две задачи: 1) выбирается тип функции f_α и 2) определяется вектор параметров (коэффициентов) A. Поиск искомой эмпирической закономерности обычно идет путем минимизации выбранного функционала невязки $E\Psi[y - f_\alpha(X,A)]$ (E - знак математического ожидания) для разных видов функций f_α. Под "разными функциями" здесь понимаются разные аналитические выражения функции $f_\alpha(X,A)$, разные преобразования и трансформации признаков, а также их подмножества.

Информацией для решения указанных двух задач служит выборка наблюдений $y_1X_1, \ldots, y_NX_N$. При ограниченном объеме выборки N вектор параметров A эмпирической закономерности оценивается с определенной погрешностью, ввиду чего значение функционала невязки $E\Psi[y - f_\alpha(X,\hat{A})]$ (здесь $\hat{A}$ - оценка вектора параметров A) возрастает по сравнению со случаем точного определения вектора параметров A, что в принципе можно достичь, неограниченно увеличивая объем выборки. В реальных ситуациях очень часто вид функции $f(X,A)$ тоже бывает неизвестным и подбирается опытным путем. Критерием выбора

здесь служит эмпирическое значение невязки, определяемое по конкретной выборке. Ввиду ограниченного объема выборки оценка величины невязки тоже получается неточной, что ведет к общему увеличению значения невязки из-за неоптимального выбора типа функции f_α.

При поиске эмпирических закономерностей вектор параметров A в функции $f_\alpha(X,A)$ обычно находится по информации, содержащейся в выборке. Для выбора типа модели часто полностью или частично используется дополнительная априорная информация. Поэтому желательно каким-нибудь способом сравнить количество информации, необходимое для определения вектора параметров A эмпирической закономерности, с информацией, необходимой для выбора модели. В качестве меры количества информации может быть использовано значение прироста невязки (вероятности ошибочной классификации, среднеквадратического отклонения прогноза), возникающего из-за неточного обучения или выбора модели при конечном объеме выборки. Далее информационный анализ поиска эмпирических закономерностей будем проводить для одного частного случая обнаружения закономерностей - задачи построения правила классификации для случая двух классов, в которой выходная переменная y в зависимости $y = f_\alpha(X,A)$ принимает лишь два значения. В качестве невязки будем использовать критерий вероятности ошибочной классификации.

§I. Прирост ошибки классификации из-за неточного определения параметров классификатора (обучения)

В настоящее время в распознавании образов известно около 200 разных способов построения правила классификации, отличающихся подходом к построению правила, сложностью, используемыми предпосылками и т.д. (см., например, [2]). Из них в настоящем разделе рассмотрим пять классификаторов, построенных по теории статистических решений, предполагая нормальность распределений признаков. Пять разных классификаторов, отличающихся между собой сложностью, т.е. числом параметров распределений признаков, участвующих в их построении, получаем при разных предположениях о структуре ковариационных матриц [2-4].

I) Никаких предположений о структуре ковариационных матриц не делается (первый классификатор). Тогда при классификации двух классов по выборке оценивается $r = p(p+3)$ параметров распределений признаков: p средних и $p(p-1)/2$ коэффициентов ковариационных матриц каждого класса (здесь p - число признаков). Дискриминантная функция получается квадратичной.

2) Ковариационные матрицы обоих классов считаются равными (второй классификатор). Число параметров $r = p(p+5)/2$ (стандартная линейная дискриминантная функция Фишера).

3) Ковариационные матрицы обоих классов считаются равными и диагональными (признаки независимы), дискриминантная функция линейная, $r = 3p$.

4) Ковариационные матрицы обоих классов считаются пропорциональными единичной матрице, поэтому при построении правила классификации не участвуют. Дискриминантная функция линейная, $r = 2p$ (классификатор евклидова расстояния).

5) Ковариационные матрицы считаются диагональными (признаки независимые). Дискриминантная функция квадратичная, $r = 4p$.

Ввиду неодинакового числа параметров, оцениваемых по выборке, чувствительность упомянутых пяти классификаторов к объему обучающей выборки будет неодинаковой. Получены аналитические формулы для ожидаемой ошибки классификации EP_N (математического ожидания вероятности ошибки классификатора, обученного по случайно полученной выборке объемом по N реализаций с каждого класса) в зависимости от объема выборки N, числа признаков p и параметров двух многомерных нормальных распределений (см.,например, ссылки к работам [3,4]). Точные формулы сложны, ненаглядны, поэтому для качественного анализа вероятностей ошибочной классификации удобнее пользоваться приближенными, асимптотическими формулами, которые, кста - ти, справедливы лишь при большом числе признаков p и большом объеме выборки N . Их общий вид

$$EP_N = \Phi\left\{-\frac{\delta}{2}\cdot\frac{\sqrt{N}}{\sqrt{\alpha}}\right\}, \qquad (1)$$

где δ - расстояние Махаланобиса между классами,определяющее "разделенность классов" или асимптотическую вероятность ошибки $P_\infty = \Phi\{-\frac{\delta}{2}\} = \lim_{N\to\infty} EP_N$ (здесь имеется в виду сферически нормальное распределение признаков); $\Phi\{a\}$ - функция Лапласа; α - коэффициент сложности, зависящий от типа правила классификации:

для правила I $\quad \alpha = 1 + \dfrac{(\delta^2 + p)^2 + p^2}{2\delta^2(N-p)}$,

для правила 2 $\quad \alpha = (1 + \dfrac{2p}{N\delta^2})/(1 - \dfrac{p}{2N})$,

для правила 3 и 4 $\alpha = 1 + \frac{2p}{N\delta^2}$,

для правила 5 $\alpha = 1 + \frac{4p}{N\delta^2}$.

Выражения для коэффициента α показывают, что, кроме объема выборки, вероятность ошибочной классификации зависит от сложности классификатора и особенно от числа признаков p . Поэтому при фиксированном объеме выборки иногда вместо сложных лучше использовать более простые классификаторы, а часть малоинформативных признаков отбросить.

§2. Прирост ошибки классификации из-за неточного выбора модели

2.I. Выбор признаков при классификации двух нормальных совокупностей с независимыми признаками впервые рассматривался Л.Д.Мешалкиным [5]. Для произвольных распределений аналитически эту задачу рассматривал В.И.Сердобольский [6]. Некоторые численные результаты о влиянии ограниченности выборки на прирост вероятности ошибки при выборе признаков приведены в работе [7]. В ней рассматривались нормально распределенные независимые признаки с одинаковыми дисперсиями σ^2 , из p исходных выбирались p_1 "лучших" по критерию разности средних $|\bar{X}_{1j} - \bar{X}_{2j}|$, $j = 1,2,\ldots,p$. Получено и проверено моделированием приближенное асимптотическое выражение для вероятности ошибочной классификации

$$P = \Phi\left\{ -\frac{\delta}{2} \cdot \frac{1}{\sqrt{1 + \frac{2p_1}{N\delta^2} \cdot \frac{\delta_2^2}{\delta_1^2}}} \right\} , \qquad (2)$$

где δ_1^2 - среднее расстояние Махаланобиса одного из исходных признаков (при выводе формулы (2) предполагалось, что расстояние Махаланобиса каждого из p исходных признаков $(\mu_{1i} - \mu_{2i})\sigma^{-1}$, μ_{1i}, μ_{2i}- средние i-го признака, σ^2 - дисперсия - случайная величина, подчиненная нормальному распределению с дисперсией δ_1^4); δ_2^2 - среднее расстояние Махаланобиса одного из p_1 отобранных признаков при идеальном их отборе.

При выборе признаков качество отобранных признаков δ_2^2 выше качества исходной системы признаков δ_1^2. Поэтому вероятность ошибки (2) превышает вероятность ошибочной классификации (I) без выбора признаков. При этом чем меньше признаков оставляется, тем больше относительный прирост вероятности ошибочной классификации.

2.2. Выбор наилучшего варианта распознающего алгоритма по критерию вероятности ошибочной классификации. При решении реальных задач признаки редко можно считать независимыми. Поэтому критерий выбора признаков должен учитывать качества не индивидуальных признаков, а всего их набора. Одним из основных критериев качества признаков или алгоритма классификации является критерий вероятности ошибочной классификации. Из известных методов оценки вероятности ошибки наиболее предпочтительны непараметрические [8]: метод "скользящего экзамена" или подсчет частоты ошибок на независимой контрольной выборке.

В работе [9] была рассмотрена задача выбора наилучшего варианта распознающего алгоритма (системы признаков, типа правила классификации, методов выделения и преобразования признаков), когда из M возможных вариантов случайным образом выбирается m вариантов ($m \ll M$), проводится оценка вероятности ошибки выбранных m вариантов и из полученных оценок $\hat{P}_1, \ldots, \hat{P}_m$ выбирается наименьшее. Здесь возможны три вида вероятностей ошибок:

$\hat{P}_{ekstr}(m)$ - экстремальное значение оценок $\hat{P}_1, \ldots, \hat{P}_m$, которое, на первый взгляд, казалось бы, и является вероятностью ошибки наилучшего варианта; назовем его кажущейся вероятностью ошибки классификации;

$P_{ideal}(m)$ - экстремальное значение вероятности из m истинных значений сравниваемых вариантов $P_1, \ldots, P_m$, которую мы получили бы в идеальном случае, когда выбор наилучшего варианта проводим не по оценкам, а по точным значениям качества каждого варианта; назовем ее "идеальной" вероятностью ошибочной классификации;

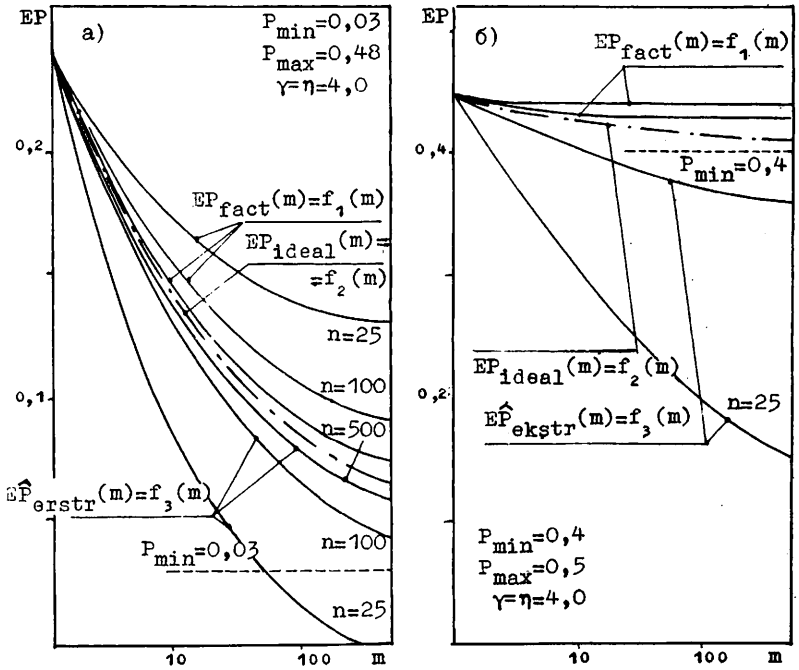
$P_{fakt}(m)$ - фактическое значение вероятности ошибки P_s, выбранного по оценкам $\hat{P}_1, \ldots, \hat{P}_m$ варианта s, для которого оценка $\hat{P}_s$ наименьшая ($\hat{P}_s \leq \hat{P}_j$, $j = 1, \ldots, m$).

Очевидно, что $P_{fakt}(m) \geq P_{ideal}(m)$, а $\hat{P}_{ekstr}(m)$ может как превышать $P_{ideal}(m)$, так и быть ниже ее. Из-за неточности оценок $\hat{P}_1, \ldots, \hat{P}_m$ разброс значений $\hat{P}_1, \ldots, \hat{P}_m$ больше, чем значений

$P_1, \ldots, P_m$, поэтому в случае, когда оценки $\hat{P}_1, \ldots, \hat{P}_m$ являются несмещенными оценками $P_1, \ldots, P_m$, можно ожидать, что математическое ожидание $E\hat{P}_{ekstr}(m) < EP_{ideal}(m)$.

В предположениях, что истинные значения качества $P_1, \ldots, P_M$ принимают v дискретных значений с априорными вероятностями $P\{P_j = \theta_\alpha\}$, $\alpha = 1, \ldots, v$, определяемыми обобщенным β-распределением в интервале P_{min}, P_{max} с параметрами формы γ и η, а условные вероятности $P\{\hat{P}_j = \frac{l}{n} | P_j = \theta_\alpha\}$ ($l = 0\,1, \ldots,$ n, n – объем выборки) – по биномиальному распределению, в работе [9] выведены формулы, приводятся подсчитанные по ним таблицы и графики для математических ожиданий вышеупомянутых трех вероятностей ошибочной классификации, в зависимости от объема выборки n, числа сравниваемых вариантов m и параметров априорного распределения Пирсона I (границ P_{min}, P_{max} и параметров формы γ и η). Здесь априорное распределение характеризует "природу" сравниваемых вариантов, и предполагается, что это распределение от экспериментатора, проводящего выбор наилучшего из m вариантов, не зависит. На рисунке приведены графики $EP_{fakt}(m) = f_1(m)$, $EP_{ideal}(m) = f_2(m)$ и $E\hat{P}_{ekstr}(m) = f_3(m)$ для трех значений объема выборки, n = 25, 100, 500, двух наборов параметров априорного распределения: 1) $P_{min} = 0.03$, $P_{max} = 0.48$; 2) $P_{min} = 0.4$, $P_{max} = 0.5$ (все варианты "плохи") при $\gamma = \eta = 4.0$ (симметричное априорное распределение). Несмотря на то, что все три вероятности ошибочных классификаций уменьшаются с ростом числа сравниваемых вариантов, поведение каждой из них индивидуально и зависит от объема выборки. При малом объеме выборки кажущаяся вероятность ошибки быстро падает (даже для случая, когда качество лучших вариантов $P_{min} = 0.4$), в то время как истинная вероятность ошибки сначала падает, достигает предела и при дальнейшем увеличении остается постоянной. При большом объеме выборки и истинная $EP_{fakt}(m)$, и кажущаяся $E\hat{P}_{ekstr}(m)$ вероятности ошибок приближаются к идеальной кривой $EP_{ideal}(m) = f_2(m)$, получаемой в случае, когда выбор проводится по точным значениям качества каждого конкурирующего варианта.

Разницу $\Delta_{выб.} = EP_{fakt}(m) - EP_{ideal}(m)$ можно рассматривать как относительное увеличение вероятности ошибки из-за недостаточно точного выбора лучшего варианта при конечном объеме выборки. Разница $\Delta_{выб.}$ зависит от объема выборки, по которой проводится оценка качества каждого конкурирующего варианта, числа сравниваемых вариантов m, и – через параметры априорного распределения (P_{min}, P_{max}, γ, η) – от самих сравниваемых вариантов.



Зависимость фактической $EP_{fakt}(m)$, "идеальной" $EP_{ideal}(m)$ и "кажущейся" $\hat{EP}_{ekstr}(m)$ вероятностей ошибок от числа сравниваемых вариантов m в зависимости от объема выборки n .

§3. Сравнение прироста ошибки классификации

Формула (2), выражающая вероятность ошибки классификатора евклидова расстояния в случае выбора p_1 признаков из p исходных, отличается от вероятности ошибки (I) с параметром $\alpha = 1 + \frac{2p}{N\delta^2}$, выражающей вероятность ошибочной классификации без выбора признаков, лишь коэффициентом $\frac{\delta_2^2}{\delta_1^2}$, где δ_2^2 - среднее качество (расстояние Махаланобиса) одного из p_1 идеальным образом отбираемых признаков, δ_1^2 - среднее качество одного из p исходных признаков. Если отбрасывается лишь малая часть признаков, то δ_2^2 мало превышает δ_1^2 и эффект выбора признаков мало сказывается на приросте вероятностной ошибки. Если выбирается лишь малая часть самих информативных признаков, то $\delta_2^2 >> \delta_1^2$ и прирост вероятности ошибки из-за выбора признаков $\Delta_{выб.}$ будет превышать соответствующий прирост из-за недостаточно хорошего обучения при конечном объеме выборки $\Delta_{обуч.} = EP_N - P_\infty$. Наиболее ясно это видно из табл. I [7].

Уже при выборе I/4 - I/8 части признаков прирост ошибки из-за неидеального выбора признаков $\Delta_{выб.}$ превышает прирост из-за неидеального обучения $\Delta_{обуч.}$, что свидетельствует о важности учета прироста ошибки из-за выбора наилучшей модели. Аналогичный вывод можно сделать и для случая, когда выбор модели проводится по критерию вероятности ошибочной классификации.

В табл. 2 приведены абсолютные значения прироста ошибки классификации из-за выбора при числе сравниваемых вариантов m = 500 для случаев $P_{min} = 0{,}03$, $P_{max} = 0{,}2$, $\gamma = \eta = 4{,}0$ (верхняя часть табл.2,"идеальная" вероятности ошибки при m = 500, $EP_{ideal}(m) = 0{,}1$) и $P_{min} = 0{,}1$, $P_{max} = 0{,}3$, $\gamma = \eta = 4{,}0$ (нижняя часть табл. 2, $EP_{ideal}(m) = 0{,}03$). Там же приводятся значения прироста ожидаемой ошибки классификации $\Delta_{обуч.} = EP_N - P_\infty$ из-за обучения для трех из пяти упомянутых в §I классификаторов I, 2 и 4.

Величина прироста ошибки классификации из-за выбора $\Delta_{выб.}$ в табл.2 приведена для 500 сравниваемых вариантов. Как видно из рисунка, при большом числе сравниваемых вариантов $\Delta_{выб.}$ увеличивается, а при меньшем - уменьшается. Величина прироста из-за обучения $\Delta_{обуч.}$ зависит, в первую очередь, от сложности используемого правила классификации и от числа признаков p_1. Несмотря на то, что в каждом конкретном выборе модели величина $\Delta_{выб.}$ зависит от

Таблица I

Абсолютное увеличение вероятности ошибочной классификации из-за обучения и из-за отбора признаков при $P_\infty = 0{,}1$

Число выбранных признаков, p_1	Прирост ошибки из-за обучения, $\Delta_{обуч.}$	Прирост ошибки из-за отбора признаков, $\Delta_{выб.}$			Объем выборки $N_1=N_2=\frac{n}{2}$
		$\frac{p_1}{p}=1/8$	$\frac{p_1}{p}=1/4$	$\frac{p_1}{p}=1/2$	
5	0,0064	0,0139	0,0094	0,0033	5
	0,0029	0,0054	0,0041	0,0015	10
	0,0006	0,0008	0,0007	0,0003	50
20	0,0039	0,0105	0,0061	0,0027	20
	0,0018	0,0051	0,0025	0,0012	40
	0,0004	0,0008	0,0005	0,0002	200

Таблица 2

Абсолютное увеличение вероятности ошибочной классификации из-за выбора модели и из-за обучения

p_1, $EP_{ideal}(m)$	Прирост ошибки из-за выбора $\Delta_{выб.}$ при $m = 500$	Прирост ошибки из-за обучения $\Delta_{обуч.}$			Объем выборки $N_1=N_2=\frac{n}{2}$
		Классификатор евклидова расстояния (4)	Линейная дискриминантная функция (2)	Квадратичная дискриминантная функция (1)	
$p_1 = 5$, $EP_{ideal}(m)$ = 0,1	0,034	0,008	0,018	0,040	25
	0,024	0,004	0,009	0,018	50
	0,009	0,001	0,002	0,003	250
$p_1 = 20$, $EP_{ideal}(m)$ = 0,1	0,034	0,025	0,094	-	25
	0,024	0,015	0,042	0,256	50
	0,009	0,004	0,007	0,053	250
$p_1 = 20$, $EP_{ideal}(m)$ = 0,03	0,040	0,0084	0,047	-	25
	0,023	0,0048	0,023	0,090	50
	0,004	0,0014	0,004	0,011	250

конкретной выборки [9], однако большой разброс значений $\Delta_{выб.}$ и $\Delta_{обуч.}$, как отчетливо видно из табл.2, свидетельствует, что прирост вероятности ошибки из-за выбора модели соизмерим с приростом из-за обучения и даже его превышает.

Величина прироста вероятности ошибки, возникающего из-за ограниченности выборки, может служить мерой информации, извлекаемой из выборки для построения правила классификации. Величина прироста позволяет по этому критерию сравнивать количество информации, используемой как для определения параметров классификатора, так и для выбора модели построения самого классификатора.

Заключение

В работе проведено сравнение количества информации, используемой из выборки для определения параметров правила классификации, с количеством информации, используемой для выбора модели при построении самого решающего правила. Само количество информации измеряется величиной прироста вероятности ошибки классификации, возникающего из-за ограниченности выборки. При увеличении объема выборки прирост вероятности ошибочной классификации из-за выбора модели и из-за обучения уменьшается, приближаясь к нулю. При выборках небольшого объема прирост вероятности ошибки из-за выбора модели очень большой. Он соизмерим и часто даже превышает прирост из-за обучения, откуда следует, что выбор модели для построения правила классификации (метода выделения или преобразования признаков) в общем-то требует больше информации, чем построение классификатора при выбранной модели. Это также свидетельствует о том, что в распознавании образов задачи выбора типа правила классификации и системы признаков являются более важными, чем задача построения самого классификатора.

Большое количество информации, необходимое для выбора модели при построении правила классификации, по-видимому, свойственно не только задаче распознавания образов, в которой выходная переменная y в зависимости от функции "вход-выход $y = f(x)$" измеряется в шкале наименование, но и в общем случае поиска эмпирических закономерностей. Полученный вывод свидетельствует, что при поиске эмпирических закономерностей, особенно в стадии выбора модели и

прогнозирующих переменных, нужно брать большие выборки либо использовать источники дополнительной информации: данные анализа физики исследуемого явления, механизма порождения измеряемых признаков, опыт решения других аналогичных задач. Нужно также использовать новые постановки задач с неоднородными данными. Поясним это более детально.

У многих экспериментаторов имеются данные, когда для одной части объектов измерен один набор признаков, для другой части - другой набор и т.д. При этом в каждой части метод и точность измерения отдельных признаков, их интерпретация могут меняться. Примером таких данных могут служить данные, собранные в разных организациях, разными людьми, в разные годы. Большинство существующих методов обработки таблиц экспериментальных данных и поиска в них закономерностей позволяют обработать таблицы, в которых для всех объектов измерены перечисленные выше признаки, ввиду чего для анализа может быть использована лишь часть выборки. При дальнейшем развитии методов обработки данных нужно разработать также методы, которые позволили бы использовать все подобного рода таблицы данных и тем самым увеличить объем выборки.

Другим типом неоднородных данных являются такие, в которых для всех частей объектов измеряются те же самые признаки, но объекты в каждой части измерены при разных условиях. Например, больные, изученные в разных клиниках; параметры изделий, изготовленных при разных технологиях и т.д. Поиск закономерности по всем данным (определение параметров правила классификации или уравнения регрессии), ввиду неоднородности данных, может привести к неудовлетворительным результатам, однако здесь все данные наверняка могут быть использованы для решения другой задачи, требующей больше информации, чем для обучения, например, выбора модели и эффективной системы признаков. Здесь правило классификации (регрессии) будет строиться по отдельным однородным выборкам, а выбор модели и признаков - по всей выборке. Таким образом, переходим от решения одной задачи построения алгоритма классификации или уравнения регрессии к одновременному решению группы задач, тем самым позволив для выбора модели, требующего много информации, использовать всю выборку, а для построения классификаторов по одной выбранной модели, оставив малые, однородные подвыборки.

Литература

I. ЗАГОРУЙКО Н.Г. Некоторые проблемы автоматического обнаружения закономерностей. -В кн.: Машинные методы обнаружения закономерностей. Новосибирск, 1976, с.6-15.

2. РАУДИС Ш. Алгоритмы построения правила классификации. Обзор. -В кн.: Статистические проблемы управления, 1975, вып. 11, с.11-51. (Ин-т матем. и киберн. АН Лит ССР).

3. РАУДИС Ш. Ограниченность выборки в задачах классификации. - В кн.: Статистические проблемы управления, 1976, вып. 18, с. 2-183. (Ин-т матем. и киберн. АН ЛитССР).

4. RAUDYS Š., PIKELIS V. On dimensionality, sample size, classification error and complexity of classification algorithm in pattern recognition. - IEEE Trans.on Pattern Analysis and Mach. Intell.,1980,v.1,PAMI-2,N 3,p.242-252.

5. МЕШАЛКИН Л.Д. Теория статистического исследования хронически протекающих болезней: Дис. на соиск. учен. степ. д-ра физ.-мат. наук. - М., Б.и.,1977. - 293 с.

6. СЕРДОБОЛЬСКИЙ В.И. Дискриминантный анализ при большом числе переменных. -Докл. АН СССР, 1980, т.254, №1, с.39-44.

7. РАУДИС Ш. Ошибки классификации при выборе признаков. -В кн.: Статистические проблемы управления, Вильнюс,1979, вып.38,с.9-26 (Ин-т матем. и киберн. АН ЛитССР).

8. RAUDYS Š. Comparison of the estimates of the probability of missclassification.- Proc.IV Intern.Joint Conf. on Pattern Recognition, Kyoto (Japan),1978,p.280-282.

9. РАУДИС Ш. Влияние объема выборки на точность выбора модели в задаче распознавания образов. -В кн.: Статистические проблемы управления, Вильнюс, 1981, вып. 50, с. 9-27 (Ин-т матем. и киберн. АН ЛитССР).

Поступила в ред.-изд.отд.
28 февраля 1981 года