КЛАССИФИКАЦИЯ БОЛЬШИХ МАССИВОВ ДАННЫХ В УСЛОВИЯХ МАЛОЙ АПРИОРНОЙ ИНФОРМАЦИИ

И. А. Пестунов, Д. И. Добротворский, Ю. Н. Синявский
Институт вычислительных технологий, Новосибирск, Россия
e-mail: pestunov@ict.nsc.ru

An effective feature extraction method for processing of large data sets and a hierarchical nonparametric classifier, developed on its basis, are presented. Results of experimental study on the model and real data confirming its efficiency are presented.

Введение

При решении прикладных задач классификации зачастую отсутствует какая-либо априорная информация о вероятностных характеристиках классов. В этих условиях целесообразно использовать алгоритмы распознавания на основе непараметрических оценок плотности распределения [1]. Они не требуют "жестких" ограничений на вид условных плотностей распределения (унимодальности, нормальности и т. п.), обеспечивая при этом высокое качество классификации. Основной недостаток таких алгоритмов — высокая вычислительная сложность. Даже несмотря на значительно возросшие за последние годы возможности вычислительной техники, применение большинства традиционных непараметрических алгоритмов классификации непосредственно для обработки больших массивов данных приводит к неприемлемым вычислительным затратам.

Один из способов решения этой проблемы — снижение размерности пространства признаков. К настоящему времени проблема извлечения информативных признаков в рамках параметрического подхода хорошо изучена, предложен ряд эффективных методов ее решения [2]. Однако практически отсутствуют методы выбора признаков для непараметрических классификаторов [3, 4]. Предложенный в работе [3] метод извлечения информативных признаков является эффективным в смысле вероятности ошибки классификации. Он обеспечивает хорошие результаты в двухклассовом случае, однако с ростом числа классов информативность выделяемых признаков существенно снижается. Кроме того, он является вычислительно сложным.

В данной статье представлены модификация метода, предложенного в [3], и разработанный на ее основе иерархический непараметрический классификатор, использующий оценки Розенблатта — Парзена.

© Институт вычислительных технологий Сибирского отделения Российской академии наук, 2007.

1. Предлагаемая модификация метода извлечения признаков

Традиционный подход к построению непараметрических правил классификации, основанных на оценках Розенблатта — Парзена, заключается в подстановке в байесовское решающее правило вместо неизвестных вероятностных характеристик классов соответствующих им оценок, полученных по обучающим выборкам [1]. Общий вид этих правил для $(0,1)$-матрицы потерь можно представить выражением

$$\hat{\delta}_0 = \hat{\delta}_0(x;V) = \arg\max_{i\in\{1,\ldots,M\}} q_i \hat{f}_i(x).$$

Здесь q_i — априорная вероятность i-го класса, $i = 1,\ldots,M$; $x \in \mathbb{R}^k$; $V = \bigcup_{i=1}^{M} V_i$ — обучающая выборка объема $N = \sum_{i=1}^{M} N_i$, $V_i = \left\{x_j^{(i)} \in \mathbb{R}^k \text{ — наблюдение из } i\text{-го класса}\right\}$; $\hat{f}_i(x)$ — оценка условной плотности распределения i-го класса $f_i(x)$ в точке $x \in \mathbb{R}^k$, определяемая выражением

$$\hat{f}_i(x) = \frac{1}{N_i h^k} \sum_{j=1}^{N_i} \Phi\left(\frac{x - x_j^{(i)}}{h}\right),$$

где h — параметр сглаживания; Φ — ядро. Для случая двух классов Ω_1 и Ω_2 это правило можно переписать следующим образом:

$$\begin{cases} x \in \Omega_1, & \text{если } \hat{g}(x) = -\ln\dfrac{\hat{f}_1(x)}{\hat{f}_2(x)} < t, \\ x \in \Omega_2 & \text{— в противном случае,} \end{cases}$$

где $\hat{g}(x)$ — непараметрическая оценка функции $g(x) = -\ln(f_1(x)/f_2(x))$, а $t = \ln(q_1/q_2)$ — решающий порог.

В соответствии с методом [3] выделение информативных признаков сводится к нахождению матрицы признаков решающей границы Σ_{DBFM} и вычислению ее собственных векторов $(v_1,\ldots,v_k)$, задающих ортонормированный базис пространства признаков. Бóльшим собственным значениям соответствуют более информативные признаки.

Пусть $n(x)$ — единичный вектор нормали к решающей границе S в точке x. Тогда матрица Σ_{DBFM} определяется следующим образом:

$$\Sigma_{\text{DBFM}} = \int_S n(x)n^{\mathrm{T}}(x)f(x)dx \Big/ \int_S f(x)dx,$$

где $f(x)$ — плотность распределения вектора признаков в точке x.

Нахождение поверхности S и нормалей к ней осуществляется следующим образом. Пусть точки $x^{(1)}$ и $x^{(2)}$ правильно классифицированы и относятся к разным классам. Тогда отрезок, соединяющий эти точки, должен пересекать решающую границу. Поэтому, двигаясь вдоль этого отрезка, можно найти точку $x \in S$ с некоторой заданной точностью.

Уравнение байесовской решающей границы можно записать в виде $g(x) = t$. Тогда вектор нормали к S в точке x выражается следующим образом:

$$\bigtriangledown g(x) = \frac{\partial g}{\partial x_1}x_1 + \frac{\partial g}{\partial x_2}x_2 + \cdots + \frac{\partial g}{\partial x_k}x_k.$$

Поскольку функция $g(x)$ в непараметрическом случае неизвестна, для оценки ее градиента используется выражение

$$\bigtriangledown g(x) \approx \frac{\Delta \hat{g}}{\Delta x_1}x_1 + \frac{\Delta \hat{g}}{\Delta x_2}x_2 + \cdots + \frac{\Delta \hat{g}}{\Delta x_k}x_k.$$

Предложенная в работе [3] процедура поиска информативных признаков для двуклассового случая может быть записана в виде последовательности шагов.

Шаг 1. Классифицируем все элементы обучающей выборки.

Шаг 2. Для каждой правильно классифицированной точки $x_i^{(1)} \in V_1$ $(i = 1, \ldots, N_1)$ находим ближайшую правильно классифицированную точку $x_{j_i}^{(2)} \in V_2$ и для каждой точки $x_j^{(2)} \in V_2$ $(j \in 1, \ldots, N_2)$ находим ближайшую правильно классифицированную точку $x_{i_j}^{(1)} \in V_1$.

Шаг 3. Двигаясь вдоль отрезков, соединяющих найденные пары точек, оцениваем точки решающей границы S.

Шаг 4. Для каждой точки z_i, найденной на предыдущем шаге, оцениваем вектор нормали к решающей границе $n(z_i)$.

Шаг 5. Оцениваем матрицу признаков решающей границы

$$\Sigma_{\mathrm{DBFM}} \approx \frac{1}{L}\sum_i n(x_i)\, n^{\mathrm{T}}(x_i), \tag{1}$$

где L — число правильно классифицированных точек обучающей выборки.

Шаг 6. Определяем собственные векторы и собственные значения матрицы Σ_{DBFM}. Собственные векторы, соответствующие бóльшим собственным значениям, определяют более информативные признаки.

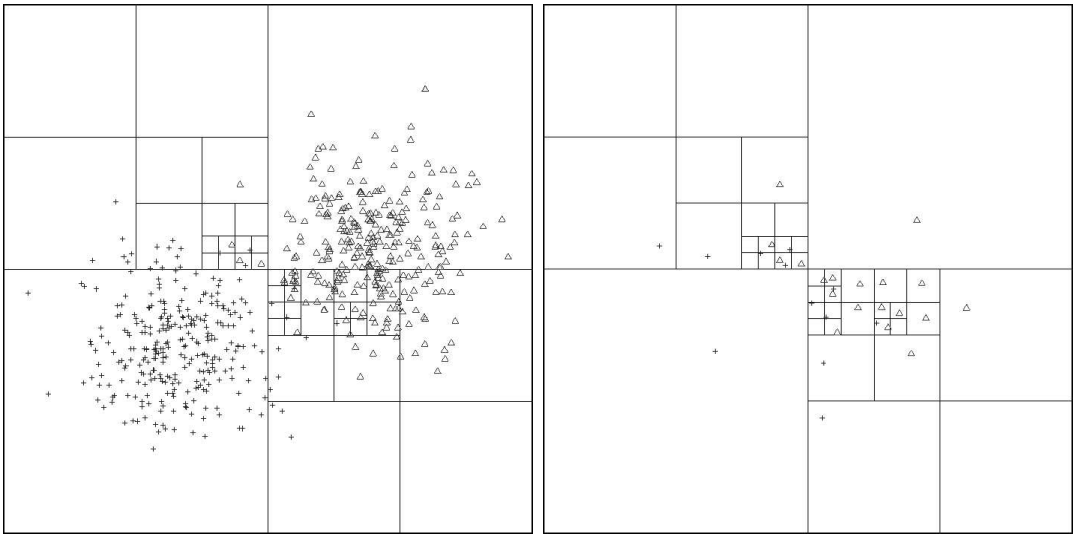
Записанный выше алгоритм имеет высокую вычислительную сложность, что существенно ограничивает его применимость к обучающим выборкам большого объема. Ниже представлен быстрый алгоритм оценивания матрицы Σ_{DBFM}. Суть этого алгоритма заключается в том, чтобы вместо исходной выборки V использовать значительно меньшую по объему рабочую выборку U, которая строится по правильно классифицированным точкам из V. Уменьшение объема выборки достигается за счет точек, удаленных от границы S, что позволяет сохранить качество оценивания решающей границы при существенном сокращении объема вычислений.

Рабочая выборка U формируется с помощью следующей рекурсивной процедуры.

Шаг 1. Находим в пространстве признаков гиперкуб, содержащий всю исходную выборку V.

Шаг 2. Если все точки выборки V, попавшие в гиперкуб, относятся к одному классу, то в U добавляем их среднее значение и останавливаем рекурсию.

Шаг 3. Если сторона гиперкуба меньше пороговой величины $\tau_{\min} = 2h$, то в U добавляем среднее значение точек каждого класса, представленного в гиперкубе, и останавливаем рекурсию.



а *б*

Рис. 1. Пример покрытия сеткой исходной (*а*) и рабочей (*б*) выборок

Т а б л и ц а 1. Результаты классификации контрольных выборок

Классификация		Число точек		
		6000	60 000	600 000
По исходной выборке	Время, с	0.821	8.201	82.819
	ВОК, %	1.4	1.836	1.793
По рабочей выборке	Время, с	0.05	0.421	4.727
	ВОК, %	1.816	1.875	1.812

Шаг 4. Делим гиперкуб пополам по каждой размерности. Полученные 2^k гиперкубов рассматриваем рекурсивно, начиная с шага 2.

При добавлении в U среднего значения некоторого набора точек новой точке u присваивается "вес" $\alpha(u)$, равный количеству этих точек. Тогда оценка (1) матрицы Σ_{DBFM} записывается в следующем виде:

$$\Sigma_{\mathrm{DBFM}} = \frac{1}{L} \sum_{i,j} n\left(u_i, u_j\right) n^{\mathrm{T}}\left(u_i, u_j\right) \alpha\left(u_i\right) \alpha\left(u_j\right).$$

В качестве иллюстрации рассмотрим работу алгоритма на двумерной модели, состоящей из двух равновероятных классов Ω_1 и Ω_2, распределенных по нормальному закону с параметрами: $\bar{\mu}^{(1)} = (0, 0)$, $\bar{\mu}^{(2)} = (2, 1)$, $\sigma_1 = \sigma_2 = (0.5, 0.5)$. Объем обучающей выборки составлял 600 точек (по 300 точек для каждого класса), а объем контрольных выборок — 6000, 60 000 и 600 000 точек. На рис. 1, *а* в графическом виде представлена исходная обучающая выборка. На ее основе построена рабочая выборка объемом 31 точка (рис. 1, *б*). Оптимальный параметр сглаживания h определен методом скользящего экзамена. При $h = 0.29$ достигнута наименьшая вероятность ошибки классификации (ВОК), равная 1.5 %. В табл. 1 приведены время и вероятность ошибки классификации контрольных выборок, полученные с использованием исходной и рабочей выборок.

Представленный алгоритм обеспечивает наиболее существенное сокращение обучающей выборки в случае, когда разделяемые классы достаточно удалены друг от друга.

2. Метод классификации для случая многих классов

Выделение информативных признаков в многоклассовом случае является достаточно сложной задачей. Для поиска признаков в случае M классов ($M > 2$) в работе [3] предлагается традиционный способ сведения ее к решению нескольких двухклассовых задач. В этом случае матрица Σ_{DBFM} определяется по формуле

$$\Sigma_{\mathrm{DBFM}} = \sum_{i>j} q_i q_j \Sigma_{\mathrm{DBFM}}\left(\Omega_i, \Omega_j\right). \tag{2}$$

Такой подход часто приводит к необоснованным вычислительным затратам и снижению качества выбираемых признаков (рис. 2).

Во многих приложениях решение общей многоклассовой задачи может быть сведено к решению нескольких задач с меньшим числом классов благодаря введению иерархии классов. Дерево классов может быть построено следующим образом.

Выберем несколько вещественных чисел r_i $(0 < r_1 < r_2 < \cdots < r_n)$, каждое из которых будет задавать i-й уровень иерархии классов. Класс i-го уровня строится из классов $(i-1)$-го уровня следующим образом.

Пусть $M^{(i)}$ — число классов i-го уровня. Тогда для всех $j \in \left\{1, \ldots, M^{(i)}\right\}$ класс

$$\Omega_j^{(i)} = \bigcup_{p \in I\left(\Omega_j^{(i)}\right)} \Omega_p^{(i-1)} \left(\Omega_j^{(0)} = \Omega_j\right),$$

где $I\left(\Omega_j^{(i)}\right)$ — множество индексов классов $(i-1)$-го уровня, образующих класс $\Omega_j^{(i)}$. При этом $I\left(\Omega_1^{(i)}\right) \cup \cdots \cup I\left(\Omega_{M^{(i)}}^{(i)}\right) = \left\{1, \ldots, M^{(i-1)}\right\}$, $I\left(\Omega_p^{(i)}\right) \cap I\left(\Omega_l^{(i)}\right) = \varnothing$ при $p \neq l$.

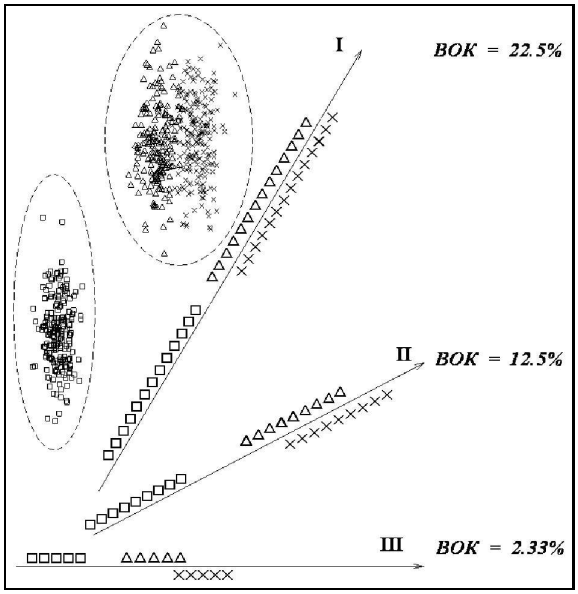


Рис. 2. Двумерная модель, состоящая из трех нормально распределенных классов: стрелками показаны оптимальный признак (III) и признаки, выделенные по методу главных компонент (I) и по формуле (2) (II)

Расстояние между классами $\Omega_p^{(i)}$ и $\Omega_q^{(i)}$ определим следующим образом:

$$\rho\left(\Omega_p^{(i)}, \Omega_q^{(i)}\right) = \min_{x \in V_p^{(i)},\ y \in V_q^{(i)}} ||x - y||,$$

где $V_j^{(i)}$ — обучающая выборка для класса $\Omega_j^{(i)}$ $(V_j^{(0)} = V_j)$; $||\cdot||$ — евклидова норма в пространстве $\mathbb{R}^k$.

Под $(n+1)$-м уровнем иерархии будем понимать объединение всех классов:

$$\Omega_1^{(n+1)} = \bigcup_{j=1}^{M^{(n)}} \Omega_j^{(n)} = \bigcup_{j=1}^{M} \Omega_j.$$

На каждом уровне иерархии для каждого класса $\Omega_j^{(i)}$ выполняются условия:

1) для любого разбиения $I\left(\Omega_j^{(i)}\right) = I' \cup I''$ существуют такие $l \in I'$ и $p \in I''$, что $\rho\left(\Omega_l^{(i)}, \Omega_p^{(i-1)}\right) \le r_i$;

2) $\rho\left(\Omega_j^{(i)}, \Omega_p^{(i)}\right) > r_i$ при $p \ne j$.

Таким образом, решение M-классовой задачи классификации сводится к решению совокупности подзадач, определяемых вершинами дерева классификации. Для каждой подзадачи строится рабочая выборка и определяется соответствующий набор информативных признаков. Процесс классификации производится последовательно с $(n+1)$-го по нулевой уровни иерархии. В каждой вершине решается локальная задача классификации с использованием соответствующих рабочей выборки и набора информативных признаков.

3. Результаты экспериментальных исследований

Как показали многочисленные экспериментальные исследования на модельных и реальных данных, введение иерархической структуры целесообразно по двум причинам:

1) решение задачи классификации при небольшом числе классов позволяет достичь высокого качества классификации с использованием всего нескольких информативных признаков;

2) расстояние между классами на ненулевых уровнях иерархии достаточно велико, поэтому соответствующие обучающие выборки значительно сокращаются.

Ниже приведены результаты нескольких экспериментов на модельных и реальных данных, подтверждающие эффективность предложенной методики анализа данных. Исследования показали, что трех уровней иерархии классов достаточно для построения эффективного решающего правила в большинстве практических задач.

Эксперимент 1. Использовалась трехмерная модель, состоящая из шести равновероятных классов $\Omega_1, \ldots, \Omega_6$, распределенных по нормальному закону с параметрами: $\bar{\mu}^{(1)} = (3, 0, 0)$, $\bar{\mu}^{(2)} = (5, 0, 0)$, $\bar{\mu}^{(3)} = (0, 3, 0)$, $\bar{\mu}^{(4)} = (0, 5, 0)$, $\bar{\mu}^{(5)} = (0, 0, 3)$, $\bar{\mu}^{(6)} = (0, 0, 5)$, $\sigma_i = (0.5, 0.5, 0.5)$, $i = 1, \ldots, 6$. Объем обучающей выборки составлял 600 точек (по 100 точек на класс), а объем контрольных выборок — 3000, 30 000 и 300 000 точек. На рис. 3 в графическом виде представлена обучающая выборка.

Для построения классификатора выделены три класса первого уровня ($\Omega_1^{(1)} = \Omega_1 \cup \Omega_2$, $\Omega_2^{(1)} = \Omega_3 \cup \Omega_4$ и $\Omega_3^{(1)} = \Omega_5 \cup \Omega_6$), а также класс второго уровня иерархии

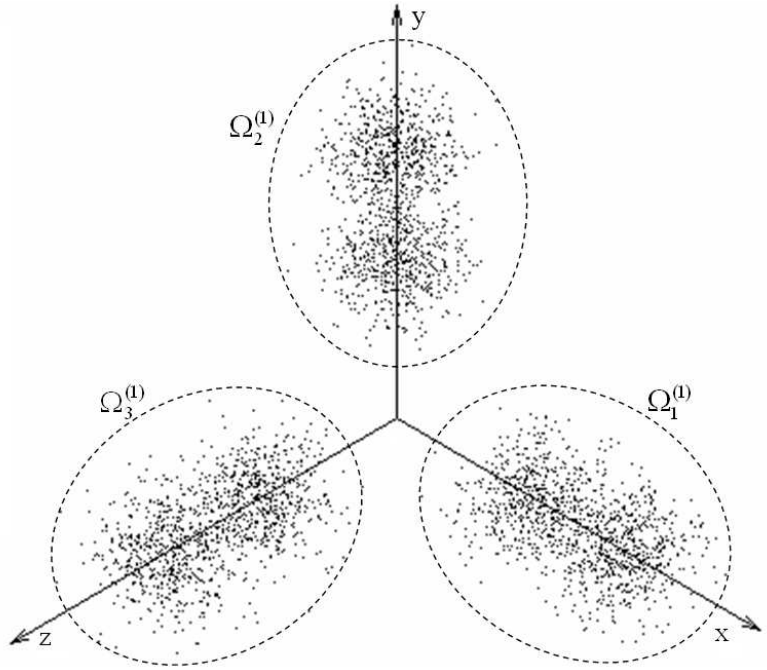


Рис. 3. Обучающая выборка. Эксперимент 1

$\Omega_1^{(2)} = \Omega_1 \cup \cdots \cup \Omega_6$. Для классов $\Omega_1^{(1)}$, $\Omega_2^{(1)}$ и $\Omega_3^{(1)}$ на основе обучающих выборок объемом 200 точек построены рабочие выборки объемом 15, 17 и 33 точки соответственно, а для класса $\Omega_1^{(2)}$ выборка уменьшилась с 600 до 12 точек.

Анализ табл. 2 показывает, что для разделения классов Ω_1 и Ω_2, Ω_3 и Ω_4, Ω_5 и Ω_6 достаточно использовать один информативный признак, а для разделения классов $\Omega_1^{(1)}$, $\Omega_2^{(1)}$ и $\Omega_3^{(1)}$ — два признака.

В табл. 3 представлены время и вероятность ошибки классификации контрольных выборок, полученные с использованием исходной и рабочей выборок.

Эксперимент 2. Эффективность алгоритма проверялась на данных, полученных Е. Андерсоном [5]. Рассматривались три класса по 50 точек в четырехмерном про-

Таблица 2. Определение достаточного для классификации числа признаков. Эксперимент 1

Число признаков	ВОК для классов, %			
	Ω_1 и Ω_2	Ω_3 и Ω_4	Ω_5 и Ω_6	$\Omega_1^{(1)}$, $\Omega_2^{(1)}$ и $\Omega_3^{(1)}$
1	0.035	0.02	0.02	0.266
2	0.035	0.025	0.025	0.0

Таблица 3. Результаты классификации контрольных выборок. Эксперимент 1

Классификатор		Объем контрольной выборки		
		3000	30 000	300 000
Байесовский	Время, с	1.72	17.22	172.32
	ВОК, %	2.84	2.73	2.65
Иерархический	Время, с	0.06	0.78	8.5
	ВОК, %	2.3	2.69	2.68

странстве признаков. При построении классификатора выделены следующие классы: $\Omega_1^{(1)} = \Omega_1$, $\Omega_2^{(1)} = \Omega_2 \cup \Omega_3$ и $\Omega_1^{(2)} = \Omega_1 \cup \Omega_2 \cup \Omega_3$. Для класса $\Omega_1^{(1)}$ объем выборки уменьшился со 100 до 13 точек, а для $\Omega_2^{(1)}$ — со 150 до 10 точек. По найденным рабочим выборкам построены следующие матрицы перехода к пространствам информативных признаков:

$$\Phi_{\Omega_2^{(1)}} = \begin{pmatrix} 0.224 & -0.956 & -0.157 & 0.103 \\ -0.061 & -0.147 & 0.957 & 0.238 \\ 0.805 & 0.096 & 0.203 & -0.549 \\ 0.545 & 0.235 & -0.126 & 0.794 \end{pmatrix}^{\mathrm{T}}, \quad \Phi_{\Omega_1^{(2)}} = \begin{pmatrix} 0.187 & 0.515 & -0.818 & 0.170 \\ -0.251 & 0.847 & 0.463 & -0.063 \\ 0.878 & 0.124 & 0.192 & -0.419 \\ 0.360 & 0.020 & 0.280 & 0.889 \end{pmatrix}^{\mathrm{T}}.$$

Вероятности ошибок классификации, полученные в ходе вычисления числа признаков, достаточного для классификации, представлены в табл. 4. Анализ таблицы показывает, что для построения классификатора достаточно использовать один информативный признак для каждой группы классов. При классификации исходной выборки иерархическим классификатором допущено четыре ошибки.

Эксперимент 3. В эксперименте использовалось изображение Краснотуранского бора (юг Красноярского края), полученное с помощью самолетного сканера С-500, разработанного в ИКИ РАН. Сканер представляет собой восьмиканальную систему. Каналам 1–8 соответствуют длины волн 800...970, 620...710, 565...630, 490...550, 525...575, 685...730, 730...800 и 900...1060 нм. Высота съемки 7300 м, размер разрешаемого элемента на местности составлял 9×18 м. Обучающая выборка, представляющая тестовые участки изображения, включала шесть классов: 1 — трава; 2 — сосновые насаждения; 3 — лиственные насаждения; 4 — поврежденные сосновые насаждения, вырубки, просеки; 5 — водная поверхность; 6 — глина. Объем обучающих выборок классов составлял 390, 570, 437, 399, 253 и 171 соответственно. Общий объем обучающей выборки 2220.

Для построения классификатора выделены четыре класса первого уровня иерархии $\left(\Omega_1^{(1)} = \Omega_1 \cup \Omega_2 \cup \Omega_3,\ \Omega_2^{(1)} = \Omega_4,\ \Omega_3^{(1)} = \Omega_5 \text{ и } \Omega_4^{(1)} = \Omega_6\right)$, а также класс второго уровня

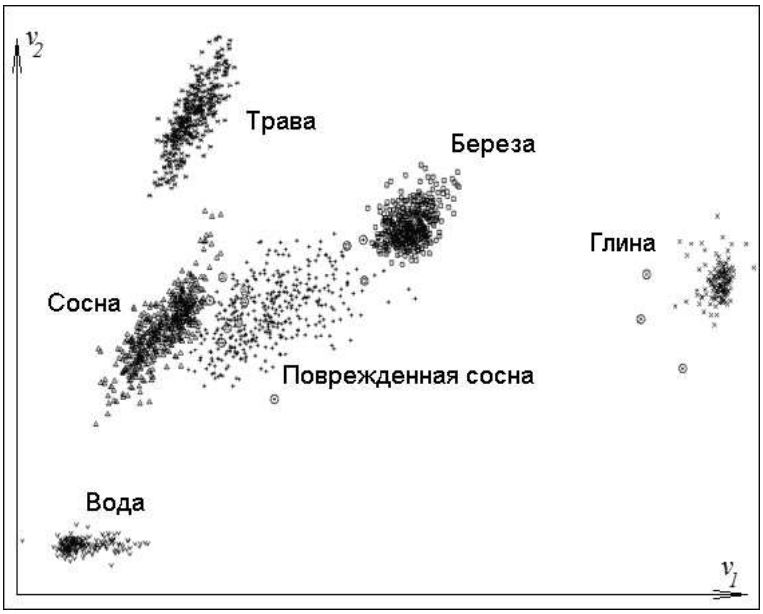


Рис. 4. Результат классификации по двум наиболее информативным признакам. Эксперимент 3

Т а б л и ц а 4. Вероятности ошибок классификации. Эксперимент 2

Число признаков	ВОК для классов, %	
	Ω_1, Ω_2 и Ω_3	$\Omega_1^{(1)}$ и $\Omega_2^{(1)}$
1	0.01	0.00
2	0.01	0.00

Т а б л и ц а 5. Вероятности ошибок классификации. Эксперимент 3

Число признаков	ВОК для классов, %	
	Ω_2 и Ω_3	$\Omega_1^{(1)}$, $\Omega_2^{(1)}$, $\Omega_3^{(1)}$ и $\Omega_4^{(1)}$
1	0.020	0.020
2	0.013	0.005
3	0.012	0.005

иерархии $\Omega_1^{(2)} = \Omega_1 \cup \cdots \cup \Omega_6$. Для класса $\Omega_1^{(1)}$ объем выборки уменьшился с 1397 до 170 точек, а для класса $\Omega_1^{(2)}$ — с 2220 до 124 точек. Вероятности ошибок классификации, полученные в ходе определения необходимого для достоверной классификации числа признаков, представлены в табл. 5. Анализ таблицы показывает, что для построения классификатора требуется не более трех информативных признаков для каждой группы классов ($\varepsilon = 0.005$).

Объем контрольной выборки составил 168 960 точек. Время работы иерархического алгоритма 23.3 с. Время работы традиционного байесовского классификатора по трем наиболее информативным признакам из исходного набора составило 276.4 с. Результаты приведены на рис. 4.

Заключение

В работе представлена модификация метода извлечения признаков, предложенного в [3], более чем на порядок сокращающая объем необходимых вычислений при незначительном снижении информативности выделяемых признаков. На ее основе построен непараметрический иерархический классификатор, позволяющий обрабатывать большие массивы данных в условиях малой априорной информации. Экспериментальные исследования подтверждают эффективность предложенного алгоритма в многоклассовом случае.

Список литературы

[1] Харин Ю.С. Робастность в статистическом распознавании образов. Минск: Университетское изд-во, 1992. 232 с.

[2] Чепонис К., Жвиренайте Д., Мирошниченко Л., Бусыгин Б. Методы, критерии и алгоритмы, используемые при преобразовании, выделении и выборе признаков в анализе данных: обзор. Вильнюс: Ин-т математики и кибернетики АН ЛитССР, 1988. 149 с.

[3] Lee C., Landgrebe D.A. Decision boundary feature extraction for non-parametric classification // IEEE Trans. on System, Man and Cybernetics. 1993. Vol. 23, N 2. C. 433–444.

[4] Abe N., Kudo M. Non-parametric classifier-independent feature selection // Pattern Recognition. 2006. Vol. 39. P. 737–746.

[5] Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. М.: Наука, 1976. С. 441–443.

Поступила в редакцию 10 сентября 2007 г.