И.К. Васильева, П.Е. Ельцов

МЕТОДЫ РАСПОЗНАВАНИЯ ОБРАЗОВ

2008

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ
Национальный аэрокосмический университет им. Н.Е. Жуковского
"Харьковский авиационный институт"

И.К. Васильева, П.Е. Ельцов

МЕТОДЫ РАСПОЗНАВАНИЯ ОБРАЗОВ

Учебное пособие по лабораторному практикуму

Харьков «ХАИ» 2008

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ
Национальный аэрокосмический университет им. Н.Е. Жуковского
"Харьковский авиационный институт"

И.К. Васильева, П.Е. Ельцов

МЕТОДЫ РАСПОЗНАВАНИЯ ОБРАЗОВ

Учебное пособие по лабораторному практикуму

Харьков «ХАИ» 2008

УДК 621.396.6

Васильева И.К. Методы распознавания образов : учеб. пособие по лаб. практикуму / И.К. Васильева, П.Е. Ельцов. – Х.: Нац. аэрокосм. ун-т "Харьк. авиац. ин-т", 2008. – 56 с.

Описаны лабораторные работы по дисциплине "Экспертные системы и распознавание образов", которая входит в программу подготовки бакалавров по направлению "Геодезия, картография и землеустройство". В текстах описаний изложены необходимые теоретические сведения, постановка задач, методика и примеры выполнения работ.

Лабораторные работы ориентированы на использование программы математических расчетов и моделирования MathCAD.

Для студентов факультета радиотехнических систем летательных аппаратов.

Ил. 21. Табл. 7. Библиогр.: 7 назв.

Рецензенты: канд. техн. наук В.И. Луценко,
канд. техн. наук С.И. Хоменко

ВВЕДЕНИЕ

Курс «Экспертные системы и распознавание образов» входит в систему подготовки специалистов широкого профиля. Реализация методов распознавания необходима в автоматизированных системах, использующих возможности искусственного интеллекта, предназначенных для решения задач диагностики, мониторинга, прогнозирования, обучения, управления поведением сложных систем в соответствии с заданными спецификациями. Такие методы теории распознавания, как кластерный анализ, выявление закономерностей в экспериментальных данных, прогнозирование различных процессов или явлений, широко используются в научных исследованиях.

Большую роль методы распознавания (классификации) играют в геоинформационных системах. Показательна в этом отношении цитата из монографии А.М. Берлянта «Геоиконика»: «...использование карт, дешифрирование снимков, анализ экранных видеоизображений, – это всегда распознавание и анализ графических образов, их измерение, преобразование, сопоставление и т.п. Отсюда следует, что распознавание графических образов, то есть создание системы решающих правил для их идентификации, классификации и интерпретации, – это одна из главных задач геоиконики».

Исторически сложилось так, что теория распознавания образов развивалась по двум направлениям: детерминистскому и статистическому, хотя чаще всего строго различить их не удается.

Детерминистский подход включает в себя математически формализованные эмпирические и эвристические методы (такие методы, в основе которых лежит моделирование процесса рассуждений). При этом используется различный математический аппарат (математическая логика, теория графов, топология, математическая лингвистика, математическое программирование и др.).

Статистический подход опирается на фундаментальные результаты математической статистики (теория оценок, последовательный анализ, стохастическая аппроксимация, теория информации).

В данном учебном пособии рассматриваются алгоритмы реализации основных типов классификаторов и методики оценки их эффективности.

Лабораторная работа № 1

ЗАДАЧА КЛАССИЧЕСКОГО ОБНАРУЖЕНИЯ. СТАТИСТИЧЕСКИЕ КРИТЕРИИ ПРИНЯТИЯ РЕШЕНИЯ

Цель работы:

– изучить методику построения решающего правила с использованием критериев максимального правдоподобия и максимума апостериорной вероятности;

– получить навыки оценивания показателей качества двухальтернативного непараметрического распознавания.

Теоретические сведения

Классификация представляет собой отнесение исследуемого объекта, задаваемого в виде совокупности наблюдений, к одному из взаимоисключающих классов. Это означает, что существует однозначное отображение совокупности наблюдений, являющейся конечным числовым множеством $\{X\}$, на множество классов $\{A\}=\{a_1,a_2,\ldots,a_k\}$, $k=1\ldots K$, $\{A\}\leftarrow\{X\}$.

В зависимости от полноты сведений о статистических характеристиках классов классификация получает наименование *различение* (при полной априорной информации о классах сигналов) или *распознавание* (при неполной априорной информации). Задача распознавания объектов в случае, когда количество классов $K = 2$, формулируется как *задача классического обнаружения*.

В классической постановке задачи распознавания универсальное множество разбивается на части – образы. Отображение объекта на воспринимающие органы распознающей системы принято называть *изображением*, а множества таких изображений, объединенные какими-либо общими свойствами, представляют собой *образы*.

Процесс, в результате которого система постепенно приобретает способность отвечать нужными реакциями на определенные совокупности внешних воздействий, называется ***обучением***. Обучение является частью процесса классификации и имеет своей конечной целью формирование эталонных описаний классов, форма которых определяется способом их использования в решающих правилах.

Методика отнесения элемента к какому-либо образу называется ***решающим правилом***. Для построения решающих правил нужна обучающая выборка. ***Обучающая выборка*** – это множество объектов, заданных значениями признаков, принадлежность которых к тому или иному классу достоверно известна "учителю" и сообщается им "обучаемой" системе.

Качество решающих правил оценивается по ***контрольной выборке***, куда входят объекты, заданные значениями признаков, принадлежность которых тому или иному образу известна только "учителю". Предъявляя объекты контрольной выборки обучаемой системе для распознавания, "учитель" может оценить качество (достоверность) распознавания.

К обучающей и контрольной выборкам предъявляются определённые требования. Например, важно, чтобы объекты контрольной выборки не входили в обучающую выборку (иногда, правда, это требование нарушается, если общий объём выборок мал и увеличить его либо невозможно, либо чрезвычайно сложно). Кроме того, обучающая и контрольная выборки должны достаточно полно представлять ***генеральную совокупность*** (гипотетическое множество всех возможных объектов каждого образа).

Основные этапы статистического распознавания – это формирование признакового пространства, получение эталонных описаний классов (если априорно эти сведения отсутствуют) и построение правила принятия решения о наблюдаемом классе объектов.

Если в результате предварительного анализа за наблюдаемой совокупностью выборочных значений возможно хотя бы приближенно установить вид закона их распределения, то априорная неопределенность относится только к параметрам этого закона; целью обучения в этом случае становится получение оценок параметров распределения. Методы распознавания, применяемые в этом случае, называются *параметрическими*. В наиболее общем случае отсутствуют априорные сведения не только о параметрах, но и о самом виде закона распределения наблюдаемой совокупности выборочных значений. Такая априорная неопределенность называется *непараметри-*

ческой, а методы распознавания, применяемые в этих условиях, – *непараметрическими*. Целью обучения в этом случае является получение оценок условных плотностей вероятностей $\hat{f}_n(\vec{x}|a_j)$.

При непараметрическом оценивании плотности вероятности используются в основном гистограммный метод, метод Парзена, метод разложения по базисным функциям, метод полигонов Смирнова, метод локального оценивания по k ближайшим соседям, а также ряд специальных методов нелинейного оценивания.

Формирование признакового пространства

Для распознавания объекты предъявляются в виде совокупности (выборки) наблюдений, обычно записываемой в виде матрицы:

$$X = \begin{matrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{p1} & x_{p2} & \ldots & x_{pn} \end{matrix}.$$

Каждый столбец $\vec{x}_i = (x_{1i}, \ldots, x_{pi})^{\mathrm{T}}$, $i = 1 \ldots n$ матрицы X представляет собой p-мерный вектор наблюдаемых значений признаков X_1, $X_2, \ldots, X_n$, являющихся безразмерными переменными.

Совокупность признаков должна в наибольшей степени отражать те свойства объектов, которые важны для классификации. При этом от размерности p признакового пространства зависит вычислительная сложность процедур обучения и принятия решения, достоверность классификации, затраты на измерение характеристик объектов.

Первоначальный набор признаков формируется из числа доступных измерению характеристик объекта $Y_1, \ldots, Y_g$, отражающих наиболее существенные для классификации свойства. На следующем этапе формируется новый набор $X_1, \ldots, X_p$; $p < g$. Традиционные способы формирования новых признаков в условиях полного априорного знания основаны на максимизации некоторой функции $J(Y_1, \ldots, Y_g)$, называемой критерием и обычно понимаемой как некоторое расстояние между классами в признаковом пространстве с координатами

$\Upsilon_1,\dots,\Upsilon_g$. В других случаях критерий $J(\Upsilon_1,\dots,\Upsilon_g)$ выражает диаметр или объем области, занимаемый классом в признаковом пространстве, и новые признаки формируются путем минимизации критерия.

Принятие решений

В теории статистических решений все виды решающих правил для $K \geq 2$ классов основаны на формировании ***отношения правдоподобия*** L и его сравнения с определенным порогом c, значение которого определяется выбранным критерием качества:

$$L = \frac{f_n(x_1, \mathbf{K}, x_n | a_2)}{f_n(x_1, \mathbf{K}, x_n | a_1)} \geq c, \qquad (1.1)$$

где $f_n(x_1, \mathbf{K}, x_n | a_j)$ – условная n-мерная плотность вероятности выборочных значений $x_1, \dots, x_n$ при условии их принадлежности к классу a_j. В статистическом распознавании эти плотности, в принципе, не известны, и в (1.1) подставляются их оценки, получаемые в процессе обучения, $\hat{f}_n(x_1, \mathbf{K}, x_n | a_j)$. Таким образом, в решающем правиле с порогом c сравнивается оценка отношения правдоподобия $\hat{L}$.

Решающее правило при использовании байесовского критерия при $K = 2$ имеет вид

$$L = \frac{f(x_1, \mathbf{K}, x_n | a_2)}{f(x_1, \mathbf{K}, x_n | a_1)} \underset{<}{\geq} \frac{\Pi_{12} - \Pi_{11}}{\Pi_{21} - \Pi_{22}} \frac{\mathrm{P}(a_1)}{\mathrm{P}(a_2)}, \qquad (1.2)$$

где $\Pi = \begin{matrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{matrix}$ – матрица потерь, элемент Π_{kl} которой количественно выражает потери от принятого решения в пользу класса a_k, когда в действительности выборка принадлежит классу a_l; $\mathrm{P}(a_j)$ – априорные вероятности классов.

Критерий (1.2) минимизирует средний риск

$$R = \sum_{k=1}^{K} \mathrm{P}(a_k) \sum_{l=1}^{K} \Pi_{kl} \mathrm{P}_{kl},$$

где P_{kl} – вероятность принятия решения о принадлежности выборки классу a_k, когда в действительности она принадлежит a_l.

Определим для $K \geq 2$ классов вероятности ошибочных решений следующим образом.

Обозначим через α_k вероятность отнесения выборки из n контрольных наблюдений к любому из классов $a_1, a_2, \ldots, a_{k-1}, a_{k+1}, \ldots, a_K$, отличному от класса a_k, когда на самом деле выборка относится именно к этому классу, а через β_k – вероятность отнесения контрольной выборки к классу a_k, когда она ему не принадлежит.

Ошибка 1-го рода – это отнесение выборки не к тому классу, к которому она принадлежит в действительности. **Ошибка 2-го рода** – это отнесение выборки к какому-либо определенному классу, к которому она на самом деле не принадлежит. При двух классах $(K = 2)$ выполняются очевидные равенства $\alpha_1 = \beta_2$ и $\alpha_2 = \beta_1$, и вероятности α_1 и β_1 совпадают с вероятностями ошибок 1-го и 2-го рода (рис. 1.1).

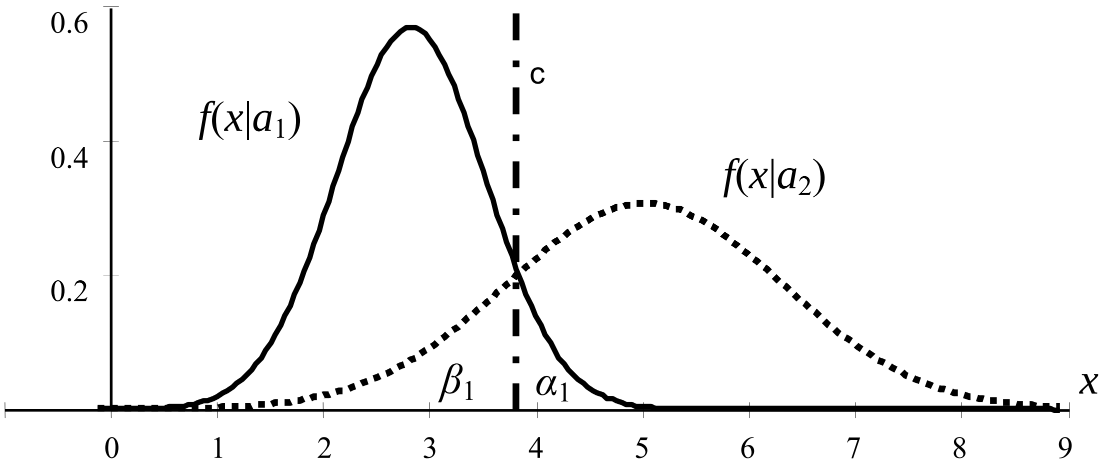


Рис. 1.1. Плотности вероятностей наблюдений классов a_1 и a_2 и порог принятия решения c

Когда матрица потерь неизвестна или ее трудно задать численно, используется критерий максимума апостериорной вероятности, согласно которому наблюдение x_n принадлежит классу a_j, чья апостериорная вероятность

$$\mathrm{P}(a_j|x_n) = \frac{\mathrm{P}(a_j) f(x_n|a_j)}{\sum_{k=1}^{K} \mathrm{P}(a_k) f(x_n|a_k)}$$

превышает апостериорные вероятности остальных классов:

$$a = \begin{cases} a_2, \text{если } L(x_n) \geq \mathrm{P}(a_1)/\mathrm{P}(a_2); \\ a_1, \text{если } L(x_n) < \mathrm{P}(a_1)/\mathrm{P}(a_2). \end{cases} \tag{1.3}$$

При использовании критериев (1.2), (1.3) в решающее правило вводятся априорные данные, но знание априорных вероятностей классов и функции потерь являются в статистическом распознавании скорее исключением, чем правилом. Кроме того, средний риск малопригоден для оценки качества классификации, когда важно в первую очередь знать ее достоверность. При отсутствии априорной информации о вероятностях состояний и потерях используются критерии Неймана–Пирсона, максимального правдоподобия и др.

Для $K = 2$ решающее правило по критерию Неймана–Пирсона имеет вид

$$a = \begin{cases} a_2, \text{если } L(x_n) \geq c; \\ a_1, \text{если } L(x_n) < c, \end{cases} \tag{1.4}$$

при этом порог c определяется таким образом, чтобы вероятность ошибочного решения P_{12} была не больше заданного значения α:

$$\mathrm{P}_{12} = \int_c^{\infty} f(L|a_1)dL \leq \alpha .$$

Использование критерия целесообразно, если одну из вероятностей ошибок можно выделить как основную и сделать ее равной некоторому требуемому значению. Однако критерий несимметричен относительно вероятностей ошибок P_{12} и P_{21}, а при классификации важно обеспечить минимальные или, по крайней мере, ограниченные заданными пределами обе вероятности ошибочных решений.

Критерий максимального правдоподобия

$$a = \begin{cases} a_2, \text{если } L(x_n) \geq 1; \\ a_1, \text{если } L(x_n) < 1 \end{cases} \tag{1.5}$$

не требует знания априорных вероятностей классов и функции потерь, позволяет оценивать достоверность решений, обобщается на случай многих классов, прост в вычислениях. Поэтому критерий (1.5) широко применяется в практических задачах распознавания образов.

Порядок выполнения работы

1. Для заданных (согласно варианту) значений параметров нормальных законов распределения $(m_1,\ \sigma_1)$ и $(m_2,\ \sigma_2)$, характеризующих два класса объектов наблюдения a_1 и a_2, определить условные по классу плотности вероятности результатов наблюдений $f(x|a_1) = f(x, m_1, \sigma_1)$ и $f(x|a_2) = f(x, m_2, \sigma_2)$.

2. Построить решающее правило по критерию максимального правдоподобия (1.5).

3. Рассчитать теоретические величины вероятностей ошибок распознавания первого и второго рода по критерию (1.5).

4. Для заданных (согласно варианту) значений априорных вероятностей $\mathrm{p}1$ и $\mathrm{p}2$ появления классов a_1 и a_2 определить условные плотности полной вероятности результатов наблюдений и апостериорные вероятности классов a_1 и a_2.

5. Построить решающее правило по критерию максимальной апостериорной вероятности (1.3).

6. Рассчитать теоретические величины вероятностей ошибок распознавания первого и второго рода по критерию (1.3).

7. Сравнить эффективности решающих правил, построенных по критериям максимального правдоподобия и максимальной апостериорной вероятности.

8. Оформить отчет о лабораторной работе, который должен содержать краткие теоретические сведения, результаты расчетов, графики исследуемых статистических характеристик и выводы.

Варианты заданий к лабораторной работе № 1

Номер варианта	m_1	σ_1	m_2	σ_2	p1	p2	Номер варианта	m_1	σ_1	m_2	σ_2	p1	p2
1	2	0.5	4	1	0.3	0.7	5	-1	0.3	1	0.9	0.6	0.4
2	0	0.4	2	1	0.4	0.6	6	-3	1	-1	0.5	0.9	0.1
3	0	1	2	0.8	0.1	0.9	7	-4	0.5	-1	1.2	0.7	0.3
4	-2	0.7	0	0.4	0.2	0.8	8	3	0.8	5	1	0.8	0.2

Для всех вариантов число точек для построения графиков $\mathrm{N} = 200$.

Контрольные вопросы

1. Что такое «отношение правдоподобия»?
2. Перечислите наиболее известные статистические критерии принятия решений. В чем их сходство и чем они отличаются?
3. Как выбрать наиболее подходящее решающее правило?
4. Как определяются вероятности ошибок первого и второго рода в задаче классического обнаружения?
5. Назовите показатели качества многоальтернативного параметрического распознавания.

Пример выполнения лабораторной работы

1. Исходные данные: число классов объектов – 2, закон распределения признаков объектов – нормальный. Параметры распределения (математическое ожидание m и среднеквадратическое отклонение σ): $m_1 = 5$, $\sigma_1 = 1$ (класс 1) и $m_2 = 3$, $\sigma_2 = 0{,}6$ (класс 2).

2. Для построения в системе MathCAD графиков условных по классу a_k $(k = 1, 2)$ плотностей вероятности признаков x

$$f(x|a_k) = \frac{1}{\sqrt{2\pi} \cdot \sigma_k} \exp \left(\frac{-(x - m_k)^2}{2\sigma_k^2} \right)$$

определим пользовательскую функцию трех аргументов:

$$f(z, m, \sigma) := \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \exp \left(\frac{-(z - m)^2}{2 \cdot \sigma^2} \right) .$$

Сформируем массив N точек ($N = 200$) по оси $0x$, располагающихся с равным шагом в диапазоне $[x_{min}, x_{max}]$. Верхнюю x_{max} и нижнюю x_{min} границы диапазона определим по правилу «трех сигм», согласно которому случайная величина x, распределенная по нормальному закону, находится в интервале значений $m \pm 3\sigma$ с вероятностью более $0{,}997$. Считаем, что случайные значения параметра x будут лежать в диапазоне $[x1min, x1max]$, если наблюдается класс 1 ($x \in a_1$), и в диапазоне $[x2min, x2max]$, если наблюдается класс 2 ($x \in a_2$), где

$$x1min = m_1 - 3{\cdot}\sigma_1,\ x1max = m_1 + 3{\cdot}\sigma_1,$$
$$x2min = m_2 - 3{\cdot}\sigma_2,\ x2max = m_2 + 3{\cdot}\sigma_2.$$

Определим нижнюю и верхнюю границы значений параметра x:

$$xmin := min\,(x1min,\ x2min);$$

$$xmax := max\,(x1max,\ x2max).$$

Для заданных данных $xmin = 0$, $xmax = 8$.

Разделим интервал $[xmin,\ xmax]$ на $(N - 1)$ часть и определим координаты точек разделения:

$$i := 0..N-1,\ x_i := xmin + \frac{xmax - xmin}{N-1}\cdot i\,.$$

Сформируем массивы значений условных по классу плотностей вероятности $f(x_i|a_1)$ и $f(x_i|a_2)$, соответствующие точкам x_i:

$$\mathrm{fx1}_i := f(x_i, m1, \sigma 1),\ \mathrm{fx2}_i := f(x_i, m2, \sigma 2).$$

Построим графики условных плотностей вероятности (рис. 1.2).

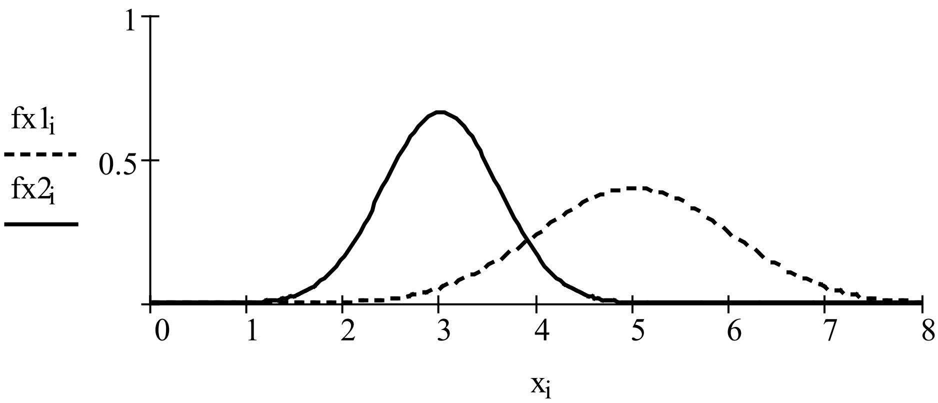


Рис. 1.2. Условные по классу плотности вероятности признака x

3. Для определения порогов принятия решения по критерию максимального правдоподобия (1.5) нужно решить уравнение

$$\frac{(x-m_1)^2}{\sigma_1^2} - \frac{(x-m_2)^2}{\sigma_2^2} = 2\ln\left(\frac{1}{\sqrt{2\pi}\cdot\sigma_1}\right) - 2\ln\left(\frac{1}{\sqrt{2\pi}\cdot\sigma_2}\right).$$

Отсюда

$$x^2\left(\sigma_2^2 - \sigma_1^2\right) + x\left(2m_2\sigma_1^2 - 2m_1\sigma_2^2\right) + m_1^2\sigma_2^2 - m_2^2\sigma_1^2 - 2\sigma_1^2\sigma_2^2\ln\left(\frac{\sigma_2}{\sigma_1}\right) = 0\,.$$

Обозначим:

$$d1 := \sigma 1^2,\ d2 := \sigma 2^2,\ a := d2 - d1,\ b := 2\cdot m2\cdot d1 - 2\cdot m1\cdot d2,$$

$$c := m1^2\cdot d2 - m2^2\cdot d1 - 2\cdot d1\cdot d2\cdot\ln\left(\frac{\sigma 2}{\sigma 1}\right).$$

Вычислим пороги принятия решения xg1 и xg2, xg1 < xg2:

$$xg1 := \frac{-b+\sqrt{b^2-4\cdot a\cdot c}}{2\cdot a},\ xg2 := \frac{-b-\sqrt{b^2-4\cdot a\cdot c}}{2\cdot a};$$

получим xg1 = –0.147 и xg2 = 3.897.

4. Изобразим на графике полученные границы раздела между классами xg1 и xg2. Если какой-либо из порогов лежит в областях маловероятных значений параметра *x* для всего множества классов $A = \{a_1, a_2\}$ (в данном случае $xg1 \notin [0, 8]$), то следует переопределить нижнюю и (или) верхнюю границы *x*:

$$xmin := if\,(\,xmin > \mathrm{xg1}\,,\,\mathrm{xg1}\,,\,xmin);$$

$$xmax := if\,(xmax < \mathrm{xg2}\,,\,\mathrm{xg2}\,,\,xmax).$$

Соответственно пересчитываются значения массивов x_i, $fx1_i$, $fx2_i$.

Для визуализации порогов принятия решения можно определить прямоугольную функцию (рис. 1.3) вида

$$\mathrm{fg}_i := if\,(\,\mathrm{xg1} < x_i < \mathrm{xg2}\,,\,0.5\,,\,0\,)$$

либо включить опцию Show Markers в диалоговом окне форматирования графика Format (закладка X–Y Axes) и ввести в поля маркеров, появившиеся на графике, имена переменных xg1 и xg2.

5. Для оценки эффективности решающего правила (1.5) рассчитаем теоретические величины вероятностей ошибок распознавания.

Вероятность отнести наблюдаемый признак к классу a_1, когда он в действительности принадлежит классу a_2:

$$\mathrm{P21} := \int_{x\min}^{xg1} f(z, m2, \sigma2)\mathrm{dz} + \int_{xg1}^{x\max} f(z, m2, \sigma2)\mathrm{dz}\,.$$

Вероятность принятия решения в пользу класса a_2, когда в действительности наблюдается класс a_1:

$$\mathrm{P12} := \int_{xg1}^{xg2} f(z, m1, \sigma1)\mathrm{dz}\,.$$

Получим P21 = 0.067 и P12 = 0.135.

Вероятность правильного распознавания определим как

$$\mathrm{P} := 1 - 0.5\cdot(\mathrm{P21}+\mathrm{P12}).$$

Получим P = 0.899.

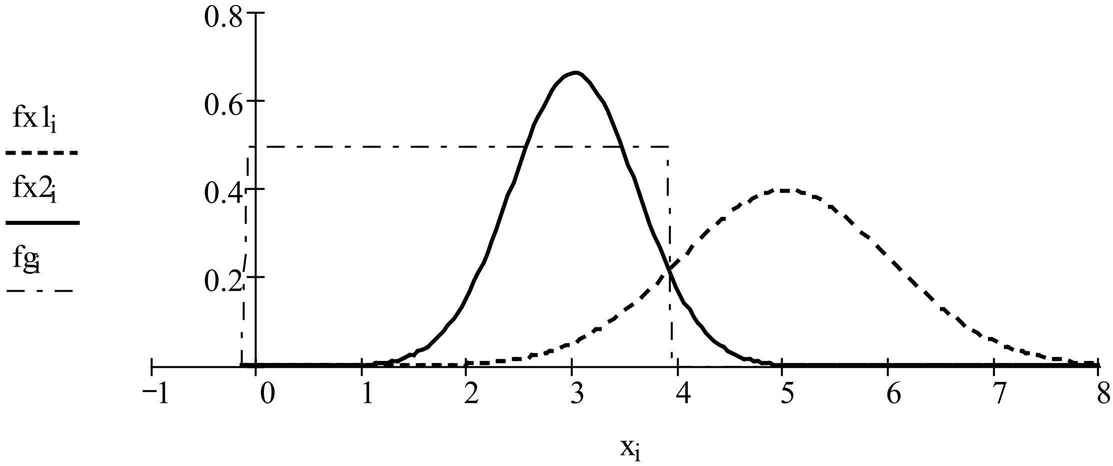


Рис. 1.3. Условные плотности вероятности fx1, fx2
и пороги принятия решения fg

6. Для построения решающего правила по критерию максимальной апостериорной вероятности (1.3) зададим априорные вероятности p1 и p2 появления классов a_1 и a_2, $\mathrm{p1} + \mathrm{p2} = 1$:

$$\mathrm{p1} := \frac{1}{3},\ \mathrm{p2} := \frac{2}{3}.$$

7. По алгоритму, описанному в п. 2, построим в интервале [$xmin$, $xmax$] график плотностей полных вероятностей появления классов a_1 и a_2 (рис. 1.4.)

$$\mathrm{fx1}_i := \mathrm{p1}\cdot f(x_i, m1, \sigma 1),\ \mathrm{fx2}_i := \mathrm{p2}\cdot f(x_i, m2, \sigma 2)$$

и график апостериорных вероятностей классов a_1 и a_2 (рис. 1.5.)

$$\mathrm{Q1}_i := \frac{\mathrm{fx1}_i}{\mathrm{fx1}_i + \mathrm{fx2}_i},\ \mathrm{Q2}_i := \frac{\mathrm{fx2}_i}{\mathrm{fx1}_i + \mathrm{fx2}_i}.$$

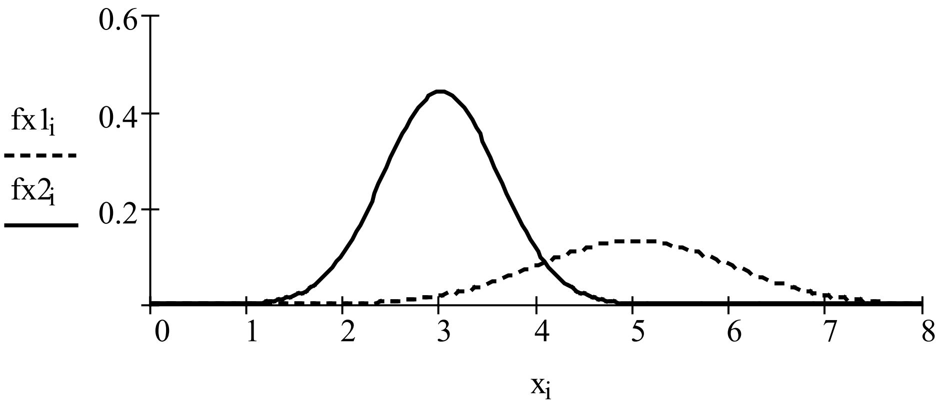


Рис. 1.4. Условные плотности полной вероятности fx1, fx2
с учетом априорных вероятностей классов

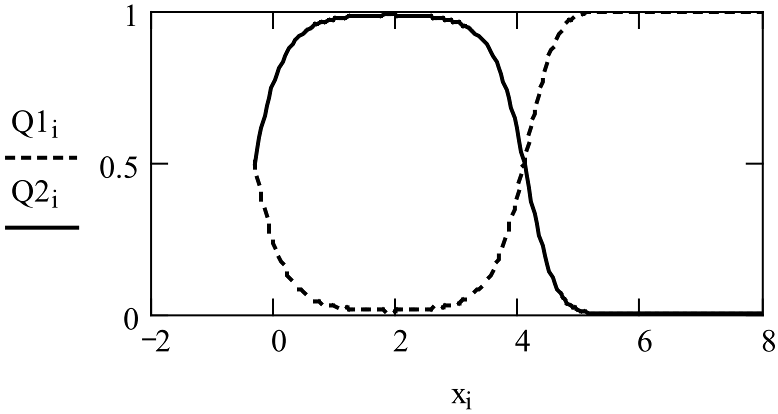


Рис. 1.5. Апостериорные вероятности Q1, Q2

8. Для определения порогов принятия решения о классе объекта по критерию максимальной апостериорной вероятности (1.3) решим квадратное уравнение

$$\frac{(x-m_1)^2}{\sigma_1^2}-\frac{(x-m_2)^2}{\sigma_2^2}=2\ln\left(\frac{1}{\sqrt{2\pi}\cdot\sigma_1}\right)-2\ln\left(\frac{1}{\sqrt{2\pi}\cdot\sigma_2}\right)-2\ln\left(\frac{\mathrm{p2}}{\mathrm{p1}}\right).$$

Приведем его к виду

$$x^2\left(\sigma_2^2-\sigma_1^2\right)+x\left(2m_2\sigma_1^2-2m_1\sigma_2^2\right)+m_1^2\sigma_2^2-m_2^2\sigma_1^2-2\sigma_1^2\sigma_2^2\ln\left(\frac{\sigma_2\mathrm{p1}}{\sigma_1\mathrm{p2}}\right)=0.$$

Обозначим:

$$d1:=\sigma1^2,\ d2:=\sigma2^2,\ a:=d2-d1,\ b:=2\cdot m2\cdot d1-2\cdot m1\cdot d2,$$

$$c:=m1^2\cdot d2-m2^2\cdot d1-2\cdot d1\cdot d2\cdot\ln\left(\frac{\sigma2\cdot\mathrm{p1}}{\sigma1\cdot\mathrm{p2}}\right).$$

Вычислим пороги принятия решения xg1 и xg2, $xg1 < xg2$:

$$xg1:=\frac{-b+\sqrt{b^2-4\cdot a\cdot c}}{2\cdot a},\ xg2:=\frac{-b-\sqrt{b^2-4\cdot a\cdot c}}{2\cdot a}.$$

Получим $xg1 = -0.332$ и $xg2 = 4.082$.

9. Для визуализации границ раздела между классами (рис. 1.6) повторим процедуру, описанную в п. 4: переопределим границы области определения признака x: $xmin$, $xmax$, пересчитаем массивы x_i, $fx1_i$, $fx2_i$, определим функцию порогов принятия решения:

$$fg_i := if\ (\ xg1 < x_i < xg2\ ,\ 0.5\ ,\ 0\ ).$$

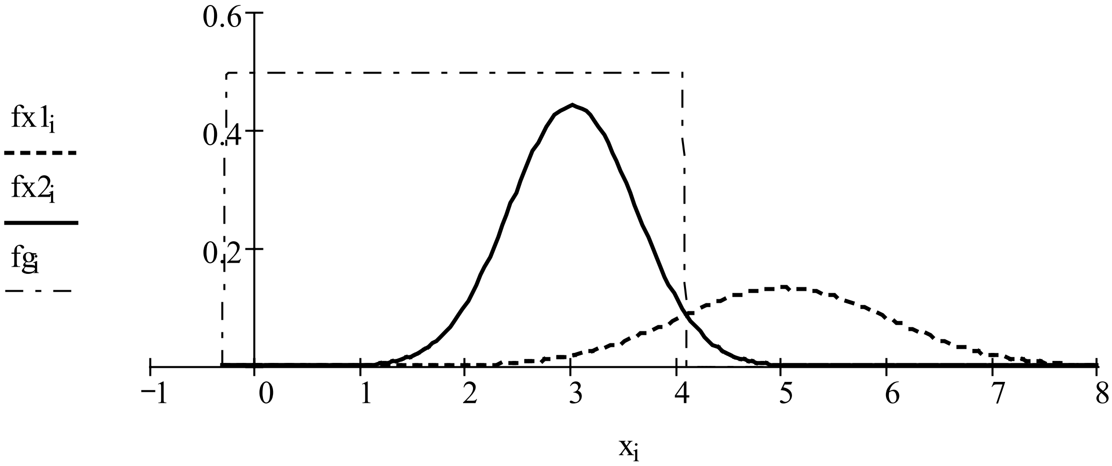


Рис. 1.6. Условные плотности полной вероятности fx1, fx2
и пороги принятия решения fg

10. Рассчитаем теоретические вероятности ошибок распознавания первого и второго рода:

$$\mathrm{P21} := \int_{x\min}^{xg1} f(z, m2, \sigma 2)\mathrm{dz} + \int_{xg1}^{x\max} f(z, m2, \sigma 2)\mathrm{dz}\,;$$

$$\mathrm{P12} := \int_{xg1}^{xg2} f(z, m1, \sigma 1)\mathrm{dz}\,.$$

Получим $\mathrm{P21} = 0.036$ и $\mathrm{P12} = 0.179$.

Вероятность правильного распознавания для случая, когда априорные вероятности классов известны и $\mathrm{p1} \neq \mathrm{p2} \neq 0.5$:

$$\mathrm{P} := 1 - (\mathrm{p1} \cdot \mathrm{P12} + \mathrm{p2} \cdot \mathrm{P21}),\ \mathrm{P} = 0.916.$$

11. На основании полученных теоретических оценок вероятностей правильного распознавания можно сделать следующие выводы:

– если сведения об априорных вероятностях классов отсутствуют, то это равносильно предположению о равных вероятностях появления классов:

$$\sum_{i=1}^{K} \mathrm{p}(a_i) = 1,\ \mathrm{p}(a_i) = \frac{1}{K},$$

где $\mathrm{p}(a_i)$ – априорная вероятность класса a_i; K – количество классов;

– случай, когда априорные вероятности классов одинаковы, является наихудшим для статистического распознавания (при прочих равных условиях).

Лабораторная работа № 2

ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ ОДНОСТУПЕНЧАТОГО АЛГОРИТМА КЛАССИФИКАЦИИ С НАКОПЛЕНИЕМ ДАННЫХ

Цель работы:

– закрепить знания о параметрических методах распознавания;

– исследовать метод распознавания одномерных нормальных совокупностей с точки зрения оптимизации временных характеристик распознающей системы;

– получить навыки реализации статистического распознавания на ЭВМ с использованием системы MathCAD.

Теоретические сведения

Классическая формулировка задачи статистического синтеза радиоэлектронных систем заключается в максимизации или минимизации статистических критериев качества системы (вероятности ошибок, времени принятия решения и т.п.) при заданных ограничениях, налагаемых на саму систему или на воздействующие на нее сигналы и процессы. Применительно к распознающим системам наибольший интерес представляет следующий вариант постановки задачи оптимизации их характеристик: минимизируется суммарное количество наблюдений (определяемое объемом обучающих и контрольной выборок и размерностью признакового пространства), необходимое для обеспечения требуемого уровня достоверности распознавания при заданном наименьшем возможном расстоянии между классами.

Достоверность решения $a = a_k$ – это вероятность принятия правильного решения P_{kk}. При количестве классов $K = 2$ и с учетом того, что $\alpha = \mathrm{P}_{12} = 1 - \mathrm{P}_{11}$ и $\beta = \mathrm{P}_{21} = 1 - \mathrm{P}_{22}$, получим для достоверностей решений a_1 и a_2 простые выражения:

$$\mathrm{P}_{11} = 1 - \alpha;\ \mathrm{P}_{22} = 1 - \beta.$$

На достоверности P_{11} и P_{22} налагается естественное ограничение

$$\mathrm{P}_{11} + \mathrm{P}_{22} > 1,$$

в противном случае можно предложить решающее правило, дающее не худшие достоверности и вообще не требующее ни обучения, ни контрольных наблюдений. Помимо этого желательно, чтобы каждая вероятность P_{11} и P_{22} была не меньше 0.5, иначе возможны случаи, когда одно из решений a_1 или a_2 лучше принимать на основе простого подбрасывания монеты – это обеспечит достоверность не хуже 0.5.

Одним из основных факторов, влияющих на достоверность классификации, является признаковое пространство $\mathrm{X}=\{x_{ij}\}:(n\times p)$, $i=1\ldots n$, $j=1\ldots p$. Считается, что значение расстояния между классами в признаковом пространстве пропорционально достоверности распознавания. Максимизация расстояния между классами (например, путем линейного преобразования $\mathbf{A}$ исходного пространства признаков Y в новое пространство X: $\mathrm{X}=\mathbf{A}\cdot\mathrm{Y}$) повышает «разделяющую силу признаков», которая, как ожидается, обеспечит требуемую достоверность различения, особенно если само решающее правило основано на том же самом критерии, что и выбор признаков.

Размерность признакового пространства p обычно стремятся сделать как можно меньше, поскольку при этом сокращается количество требуемых измерений, упрощаются вычисления, формирующие и реализующие решающие правила, повышается статистическая устойчивость результатов распознавания. Вместе с тем уменьшение p, вообще говоря, ведёт к снижению достоверности распознавания. Поэтому формирование признакового пространства является компромиссной задачей, которую можно разделить на две части: формирование исходного признакового пространства и минимизация размерности этого пространства. Формирование исходного пространства пока что основано на опыте, интуиции, а иногда и на везении. Теоретически обоснованные подходы к решению этой задачи в литературе не встречаются. В части, касающейся минимизации размерности, существуют формальные методы и алгоритмы, основанные, как правило, на исследовании корреляционных свойств признаков и последовательном исключении из признакового пространства какого-либо признака из пары наиболее коррелированных.

Возможность повышения достоверности распознавания путем увеличения количества наблюдений n, по которым принимается решение, открывает еще один путь формирования признакового пространства без применения линейного преобразования пространства исходных признаков $\mathrm{Y}_1,\ldots,\mathrm{Y}_g$. С практической точки зрения представляет интерес задача оптимизации суммарного количества наблюдений (определяемое, в общем случае, объемом обучающих и контрольной выборок и размерностью признакового пространства), необходимого для обеспечения требуемого гарантированного уровня достоверности распознавания при заданном наименьшем возможном расстоянии между классами, в качестве которого из практических соображений естественно взять реальную точность измерения этого расстояния в распознающих системах.

Заметим, что если время на принятие решений жестко ограничено, повышение размерности признакового пространства может оказаться единственным средством увеличения достоверности до требуемого уровня.

Распознавание одномерных нормальных совокупностей с общей дисперсией

Рассмотрим задачу определения принадлежности выборки, состоящей из n независимых нормально распределенных наблюдений, к одному из двух классов $A = \{a_1, a_2\}$.

Измеряемые значения признака объекта x представляют собой реализации случайной величины с плотностью распределения

$$f(x,m,\sigma)=\frac{1}{\sqrt{2\pi}\cdot\sigma}\exp\ \frac{-(x-m)^2}{2\sigma^2}\ . \qquad (2.1)$$

Для построения решающего правила (алгоритма классификации объекта) по критерию максимального правдоподобия необходимо определить точки пересечения графиков условных плотностей вероятности $f(x|a_1)=f(x,m_1,\sigma_1)$ и $f(x|a_2)=f(x,m_2,\sigma_2)$, т.е. значения признака x, для которых отношение правдоподобия равно единице:

$$L(x)=\frac{f(x,m_1,\sigma_1)}{f(x,m_2,\sigma_2)}=1. \qquad (2.2)$$

При совместной обработке совокупности значений измеряемого признака $\vec{x}=\{x_1,x_2,\ldots x_n\}$ в предположении независимости элементов x_i $(i = 1, 2,\ldots, n)$ выборки $\vec{x}$ условные по классу a_k функции правдоподобия (многомерные плотности распределения) будут иметь вид

$$f(x|a_k)=f(\vec{x},m_k,\sigma_k)=\prod_{i=1}^{n} f(x_i,m_k,\sigma_k). \qquad (2.3)$$

Логарифм отношения правдоподобия (2.2) $(\ln L(\vec{x}))$ при равных дисперсиях σ^2 функций правдоподобия (2.3) (при $m_2 > m_1$) имеет вид

$$\ln \frac{f(\vec{x},m_1,\sigma)}{f(\vec{x},m_2,\sigma)} = \frac{m_2-m_1}{\sigma^2}\cdot\sum_{i=1}^{n} x_i - \frac{n\cdot\left(m_2^2-m_1^2\right)}{2\cdot\sigma^2}. \qquad (2.4)$$

При априорно известных математических ожиданиях m_1, m_2 задача распознавания формулируется и решается в рамках теории проверки статистических гипотез как проверка простой гипотезы H_1: среднее значение нормально распределенных наблюдений равно m_1 против простой альтернативы H_2, что среднее равно m_2 при известной общей дисперсии σ. Решающая процедура заключается в сравнении логарифма отношения правдоподобия (2.4) с некоторым порогом $\ln \mathrm{c}$, зависящим от выбранного критерия качества:

$$\ln L(\vec{x}) \geq \ln \mathrm{c}\,,\ a_2 > a_1. \qquad (2.5)$$

При выполнении неравенства (2.5) принимается решение H_2, при невыполнении – H_1. Если решение принимается по критерию максимального правдоподобия (2.2), то $\ln \mathrm{c} = \ln 1 = 0$.

Поскольку логарифм – монотонное преобразование, то достаточной статистикой для принятия решений будет среднее выборочное значений

$$y_n = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i\ , \qquad (2.6)$$

и для получения решающей распознающей процедуры (2.5) можно использовать правило

$$\gamma = \begin{matrix} y_n > x_{gr} \Rightarrow \mathrm{H}_2 : A = a_2; \\ y_n \leq x_{gr} \Rightarrow \mathrm{H}_1 : A = a_1, \end{matrix} \qquad (2.7)$$

где порог принятия решения $x_{gr} = \ln \mathrm{c}$ определяется из (2.4), (2.5) как

$$\ln c = x_{gr} = \frac{m_1 + m_2}{2}. \qquad (2.8)$$

Величина $\frac{1}{n}\sum_{i=1}^{n} x_i$, представляющая собой по условию задачи сумму независимых нормально распределенных случайных величин, имеет нормальное распределение со средним

$$\mathrm{M}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \middle| a_k\right) = m_k \ (k=1,\,2)$$

и дисперсией

$$\mathrm{D}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \middle| a_k\right) = \frac{\sigma_k^2}{n} \ (k=1,\,2).$$

Тогда достоверность распознавания при использовании решающего правила (2.7) определяется через табулированный интеграл вероятности $\mathrm{F(z)}$:

$$\alpha = \beta = 1 - \mathrm{F}\left(\frac{(m_2 - m_1)\sqrt{n}}{2\sigma} + \frac{\sigma \cdot \ln c}{(m_2 - m_1)\sqrt{n}}\right), \ m_2 > m_1.$$

Последовательное принятие решений

При последовательном анализе Вальда на каждом этапе пространство выборок наблюдений разделяется на три области: допустимую G_1, критическую G_2 и промежуточную $G_{\text{пр}}$. Если выборочное значение попадает в $G_{\text{пр}}$, то делается следующее наблюдение, и так до тех пор, пока при некотором значении n_1 размера выборки выборочное значение не попадет в одну из областей (G_1 или G_2), после чего принимается одна из гипотез: наблюдается класс a_1 (при попадании в G_1) или наблюдаемая выборка принадлежит классу a_2 (G_2).

Критерием качества последовательного правила выбора решения обычно является минимум среднего значения размера выборки, необходимой для принятия решения, при заданных значениях вероятностей ложной тревоги α и пропуска сигнала β.

А. Вальдом показано, что среди всех правил выбора решения (в том числе и непоследовательных и, в частности, известных критериев – байесовского, максимума апостериорной вероятности, максимума правдоподобия, Неймана–Пирсона), для которых условные вероятности ложной тревоги и пропуска сигнала не превосходят α и β, последовательное правило выбора решения, состоящее в сравнении логарифма отношения правдоподобия $\ln L(\vec{x})$ (или отношения правдоподобия $L(\vec{x})$) с двумя порогами – нижним c_1 и верхним c_2 – приводит к наименьшим средним значениям размера выборок $M\{n|a_1\}$ (при справедливости гипотезы H_1: $a = a_1$) и $M\{n|a_2\}$ (при справедливости гипотезы H_2: $a = a_2$).

Аналитически процедура последовательного анализа может быть выражена следующим образом: при n-м наблюдении принимается гипотеза H_1, если

$$c_1 < \ln L(x_1, x_2, \ldots, x_k) < c_2,\; k = 1, 2, \ldots, n-1;\; \ln L(x_1, x_2, \ldots, x_n) \le c_1,$$

гипотеза H_2, если

$$c_1 < \ln L(x_1, x_2, \ldots, x_k) < c_2,\; k = 1, 2, \ldots, n-1;\; \ln L(x_1, x_2, \ldots, x_n) \ge c_2 .$$

Таким образом, последовательное правило выбора решения, в отличие от байесовского, предусматривает сравнение логарифма отношения правдоподобия с порогами c_1 и c_2, не зависящими от априорных вероятностей наличия или отсутствия сигнала и от потерь. Эти пороги с некоторым приближением можно выразить через заданные значения вероятностей ложной тревоги α и пропуска сигнала β:

$$c_1 = \ln[\beta/(1-\alpha)],$$

$$c_2 = \ln[(1-\beta)/\alpha].$$

Поскольку при последовательном анализе размер выборки является случайной величиной, то даже при сравнительно малых средних значениях длительности процедуры возможны случаи недопустимо больших размеров выборки. Типичным примером компромиссного решения для распределения длительности процедуры является усеченный последовательный анализ, при котором заранее устанавливается максимальное значение объема выборки $n_{\max}$, при достижении которого последовательная процедура заканчивается и соответствующее отношение правдоподобия сравнивается не с двумя порогами (c_1 и c_2), а только с одним ($c_{\text{ус}}$), в результате чего обязательно принимается одно из решений.

Порядок выполнения работы

1. Для заданных (согласно варианту) значений параметров нормальных законов распределения m_1, m_2 и σ, характеризующих два класса объектов наблюдения a_1 и a_2, определить условные плотности вероятности результатов наблюдений x: $f(x|a_1) = f(x, m_1, \sigma)$, $f(x|a_2) = f(x, m_2, \sigma)$.

2. Вычислить порог принятия решения (2.8) и формализовать решающее правило.

3. С помощью алгоритма генерации нормально распределенной случайной величины смоделировать данные наблюдений классов a_1 и a_2; размер каждого массива данных $\mathrm{N} = 100$ элементов.

4. Для исследования эффективности алгоритма распознавания с накоплением данных принять диапазон варьирования объема контрольной выборки $\mathrm{n} = 1, 2, \ldots, 10$.

5. Для текущего объема контрольной выборки вычислить N раз ($\mathrm{N} = 100$) достаточные статистики для двух классов объектов.

6. Для текущего объема контрольной выборки рассчитать теоретические вероятности ошибок распознавания первого и второго рода и найти их эмпирические оценки. В качестве показателей эффективности алгоритма распознавания принять среднее значение вероятностей ошибок и среднее значение оценок этих вероятностей.

7. Построить графики зависимостей экспериментальной и теоретической вероятностей ошибок от объема накопления данных n.

8. Оформить отчет о лабораторной работе, который должен содержать краткие теоретические сведения, алгоритмы моделирования, результаты расчетов, графики исследуемых статистических характеристик и полученных зависимостей, выводы.

Варианты заданий к лабораторной работе № 2

Номер варианта	m_1	m_2	σ
1	1	3	2
2	1	3	1.5
3	2	3	1
4	2	3	0.75

Номер варианта	m_1	m_2	σ
5	-1	1	1
6	-1	1	1.5
7	-1	0	1
8	-2	-1	0.75

Для всех вариантов число классов объектов $K = 2$, число точек для построения графиков $N = 100$, объемы выборок $n = 1, 2, 3, \ldots, 10$.

Контрольные вопросы

1. Что такое «достаточная статистика»?
2. Что называют контрольной выборкой?
3. Что общего и в чем различие задачи принятия статистических гипотез и задачи статистического распознавания?
4. Чем отличаются постановки задач обнаружения и распознавания объектов?

Пример выполнения лабораторной работы

1. Исходные данные: число классов объектов – 2, закон распределения признаков объектов – нормальный с известной общей дисперсией $D = \sigma^2 = 2$. Параметры сигналов от объектов разных классов отличаются параметрами законов распределения – математическими ожиданиями $m1 = 4$ для класса a_1 и $m2 = 6$ для класса a_2.

2. Определим пользовательскую функцию – плотность нормального распределения с параметрами m, σ:

$$f(z,m,\sigma) := \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\exp\left(\frac{-(z-m)^2}{2\cdot\sigma^2}\right).$$

По правилу «трех сигм» вычислим верхнюю $xmax$ и нижнюю $xmin$ границы области определения измеряемого параметра x:

$$Xmin := m1 - 3\cdot\sigma,\ xmax := m2 + 3\cdot\sigma.$$

Разобьем интервал [xmin, xma$x$] на (N − 1) отрезков и определим массив координат точек разбиения:

$$\mathrm{N} := 100,\ \ \mathrm{i} := 0..\mathrm{N}-1,\ \ \mathrm{x_i} := \mathrm{xmin} + \frac{\mathrm{xmax} - \mathrm{xmin}}{\mathrm{N}-1} \cdot \mathrm{i}.$$

В точках x_i вычислим значения условных плотностей распределения вероятности величины параметра x для классов a_1 и a_2:

$$\mathrm{fx1_i} := \mathrm{f}(\mathrm{x_i}, \mathrm{m1}, \sigma),\ \ \mathrm{fx2_i} := \mathrm{f}(\mathrm{x_i}, \mathrm{m1}, \sigma).$$

3. Для визуализации границы между классами определим порог принятия решения

$$\mathrm{xg} := \frac{\mathrm{m1} + \mathrm{m2}}{2}$$

и объявим вспомогательный массив

$$\mathrm{fg_i} := \mathrm{if}(\mathrm{x_i} > \mathrm{xg}, 0.35, 0)$$

(вместо объявления массива fg_i можно включить опцию Show Markers и ввести в поле маркера имя xg).

Построим графики условных плотностей вероятности параметра x с представлением границы между классами a_1 и a_2 (рис. 2.1).

4. Для моделирования результатов измерений параметра x – случайной величины с нормальным законом распределения – определим функцию Norm(m,σ):

$$\mathrm{v} := 48\ \ \ \mathrm{k} := 1..\mathrm{v}$$

$$\mathrm{Norm}(\mathrm{m}, \sigma) := \sqrt{\frac{12}{\mathrm{v}}} \cdot \sigma \cdot \ \sum_{\mathrm{k}} \mathrm{rnd}(1) - \frac{\mathrm{v}}{2}\ \ + \mathrm{m}.$$

Получим N данных наблюдений для классов a_1 и a_2:

$$\mathrm{x1_i} := \mathrm{Norm}(\mathrm{m1}, \sigma);\ \mathrm{x2_i} := \mathrm{Norm}(\mathrm{m2}, \sigma).$$

Построим графики реализаций наблюдений $x1_i$ ($x \in a_1$), $x2_i$ ($x \in a_2$) и покажем границу раздела между классами (рис. 2.2).

5. Зададим начальный объем контрольной выборки:

$$\mathrm{n} := 1.$$

6. Для текущего объема контрольной выборки рассчитаем достаточные статистики для m1 и m2:

$$j := 1..n;$$

$$y1_i := \sum_j \frac{Norm(m1, \sigma)}{n}; \qquad y2_i := \sum_j \frac{Norm(m2, \sigma)}{n}.$$

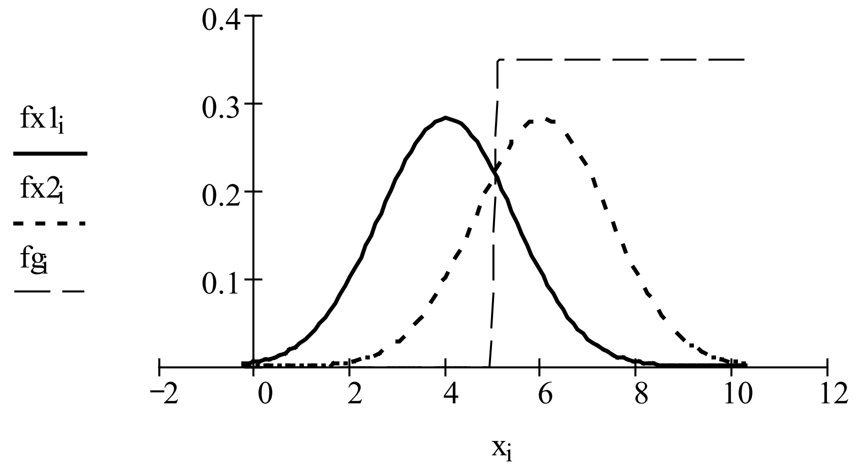


Рис. 2.1. Условные по классу плотности вероятности признака x и граница между классами

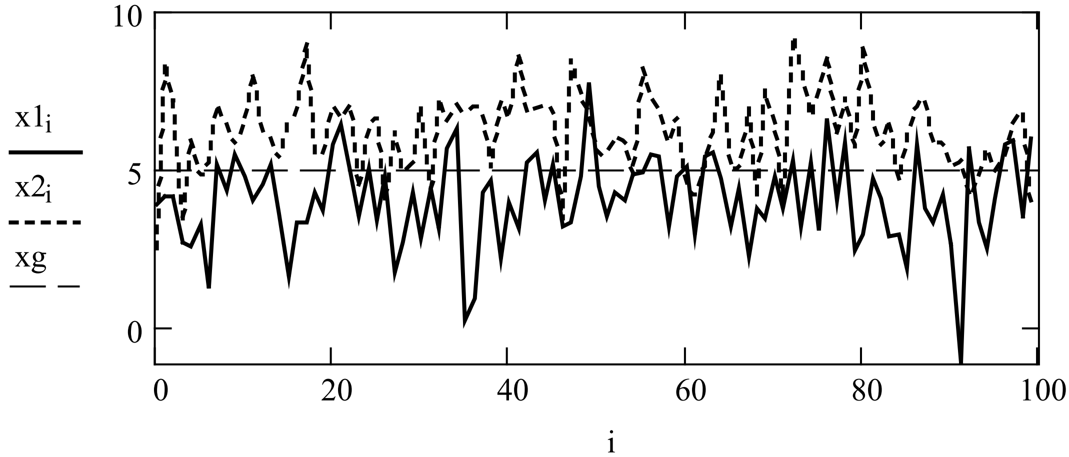


Рис. 2.2. Реализации наблюдений двух классов x1 и x2 и граница между классами xg

7. Согласно решающему правилу

$$\{\text{если } (y < xg), \text{ то } (x \in a_1), \text{ иначе } (x \in a_2)\}$$

подсчитаем количество ошибочных решений N12 (принятие решения о классе a_1, когда контрольная выборка принадлежит классу a_2) и N21 (принятие решения о классе a_2, когда контрольная выборка принадлежит классу a_1):

$$N12 := \sum_i if(y2_i < xg, 1, 0);$$

$$N21 := \sum_i if(y1_i > xg, 1, 0).$$

Рассчитаем эмпирические оценки вероятности ошибок распознавания:

$$P12\_e := \frac{N12}{N}; \qquad P21\_e := \frac{N21}{N};$$

$$P\_e := \frac{P12\_e + P21\_e}{2}.$$

8. Для анализа влияния объема n контрольной выборки на достоверность принятия решения, будем проводить расчеты для $n = 1, 2, 3, \ldots, 10$. При $n = 1$ результаты полностью соответствуют результатам распознавания без накопления информации ($y_i = x_i$). При $n = 2$ решения принимаются по совокупности двух последовательно взятых отсчетов ($y_i = (x_i + x_{i+1})/2$) и т.д.

Теоретическую вероятность ошибки распознавания можно оценить по формулам

$$P_{12} = \int_{-\infty}^{x_{gr}} f(x|a_2)dx \text{ и } P_{21} = \int_{x_{gr}}^{\infty} f(x|a_1)dx.$$

Как известно, сумма n нормально распределенных случайных чисел с математическим ожиданием m и дисперсией σ^2 распределена по нормальному закону с математическим ожиданием m и дисперсией (σ^2/n), поэтому для расчета плотностей вероятностей $f1n_i$ и $f2n_i$ статистик $y1$ и $y2$ используем пользовательскую функцию, определенную в п. 2, но с другими аргументами:

$$f1n_i := f\left(x_i, m1, \frac{\sigma}{\sqrt{n}}\right); \quad f2n_i := f\left(x_i, m2, \frac{\sigma}{\sqrt{n}}\right).$$

Вычислим теоретические вероятности ошибок:

$$P12\_t := \int_{xg}^{xmax} f\left(x, m1, \frac{\sigma}{\sqrt{n}}\right) dx;$$

$$P21\_t := \int_{xmin}^{xg} f\left(x, m2, \frac{\sigma}{\sqrt{n}}\right) dx;$$

$$P\_t := \frac{P12\_t + P21\_t}{2}.$$

9. Построим графики реализаций статистик $y1_i$ ($m = m1$) и $y2_i$ ($m = m2$) и покажем границу раздела между классами (рис. 2.3).

10. Для сравнения построим в одной координатной плоскости графики плотности вероятности реализаций наблюдений и плотности вероятности статистик (рис. 2.4).

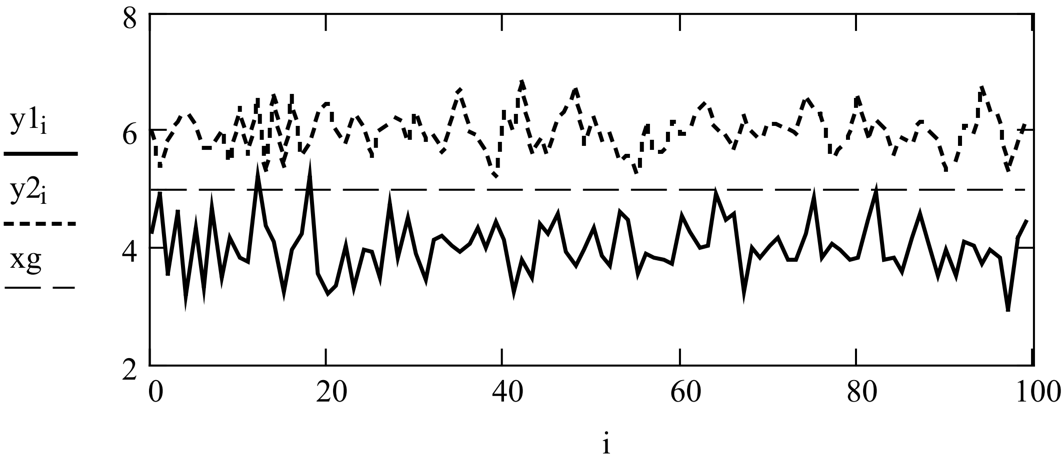


Рис. 2.3. Реализации статистик $y1$ и $y2$

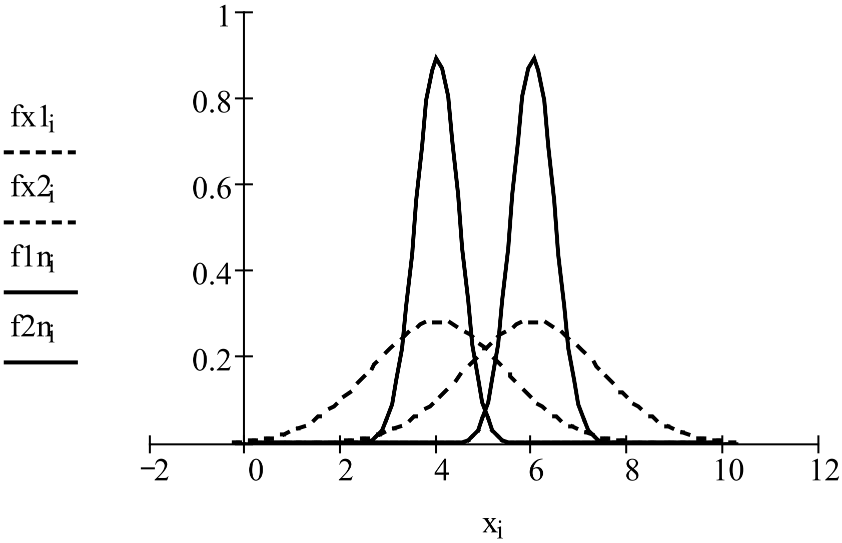


Рис. 2.4. Условные по классу плотности вероятности данных наблюдений и статистик

11. Сохраним результаты исследования эффективности классификатора (экспериментальную и теоретическую вероятности ошибки) при объеме контрольной выборки $n = 1$. Для этого выведем текущие

значения переменных P_e и P_t и присвоим элементам массивов PE_1 и PT_1 соответствующие числовые константы:

$$P\_e := 0.25; \qquad PE_1 := 0.25;$$

$$P\_t := 0.24; \qquad PT_1 := 0.24.$$

12. Повторим исследования, начиная с п. 5, для $n = 2, 3, \ldots, 10$.

Для каждого последовательно вводимого значения n $(1, 2, \ldots, 10)$ результаты расчетов P_e и P_t должны присваиваться соответствующим элементам массивов PE_j и PT_j $(j = 1, 2, \ldots, 10)$, где массивы $\{PE_j\}$ и $\{PT_j\}$ – экспериментальные и теоретические вероятности ошибки распознавания при постепенном наращивании объема данных n.

Сохраненные результаты показаны в табл. 2.1.

Таблица 2.1

n	1	2	3	4	5	6	7	8	9	10
PT_j	0.24	0.159	0.11	0.079	0.057	0.042	0.031	0.023	0.017	0.013
PE_j	0.25	0.135	0.1	0.06	0.08	0.04	0.015	0.045	0.01	0.0

12. По окончании расчетов строятся графики зависимостей экспериментальной и теоретической вероятностей ошибок от объема накопления данных n (рис. 2.5).

$$j := 1..10$$

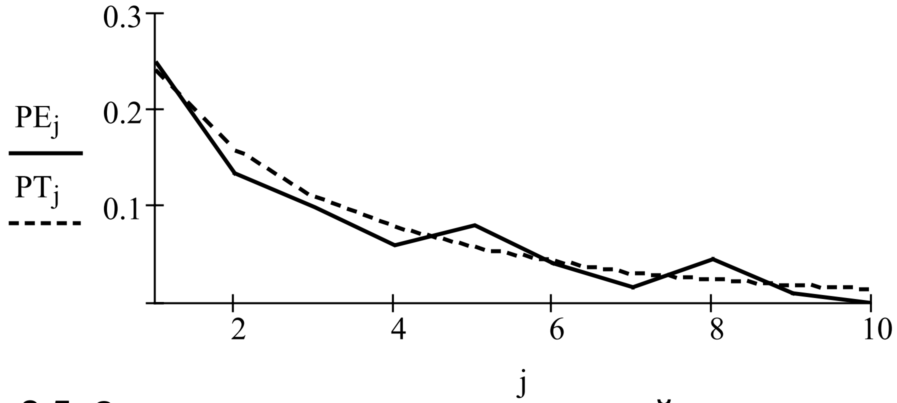


Рис. 2.5. Зависимости экспериментальной и теоретической вероятностей ошибок от объема выборки

Лабораторная работа № 3

МЕТОДЫ РАЗДЕЛЯЮЩИХ ФУНКЦИЙ

Цель работы:

– изучить непараметрические методы «обучения с учителем», основанные на линейных разделяющих функциях и методику построения кусочно-линейных решающих правил;

– получить навыки статистического оценивания показателей качества классификации с использованием системы MathCAD для моделирования и представления объектов в виде данных наблюдений.

Теоретические сведения

Обучением называют процесс выработки в некоторой системе той или иной реакции на группы внешних идентичных сигналов путем многократного воздействия на систему внешней корректировки. Механизм генерации этой корректировки определяет алгоритм обучения. Рассмотрим некоторые методы обучения с "учителем".

Метод построения эталонов

Для каждого класса по обучающей выборке строится эталон, имеющий значения признаков:

$$\vec{x}^0 = \{x_1^0, x_2^0, \ldots x_n^0\},\ x_i^0 = \frac{1}{K}\sum_{k=1}^{K} x_{ik}\ ,$$

где K – количество объектов данного образа в обучающей выборке.

По существу, эталон – это усреднённый по обучающей выборке абстрактный объект. Абстрактным его называют потому, что он может не совпадать не только ни с одним объектом обучающей выборки, но и ни с одним объектом генеральной совокупности.

Распознавание осуществляется следующим образом. На вход системы поступает объект $\vec{x}^*$, принадлежность которого к тому или иному образу не известна. От этого объекта измеряются расстояния до эталонов всех образов, и система относит $\vec{x}^*$ к тому образу, расстояние до эталона которого минимально.

Правило ближайшего соседа

Пусть $X^n = \{\vec{x}_1^*, \vec{x}_2^*, \ldots \vec{x}_n^*\}$ будет множеством n помеченных выборочных значений и $\vec{x}_i^* \in X^n$ будет точкой, ближайшей к $\vec{x}$. Тогда *правило ближайшего соседа* для классификации $\vec{x}$ заключается в том, что $\vec{x}$ присваивается метка, ассоциированная с $\vec{x}_i^*$ (рис. 3.1).

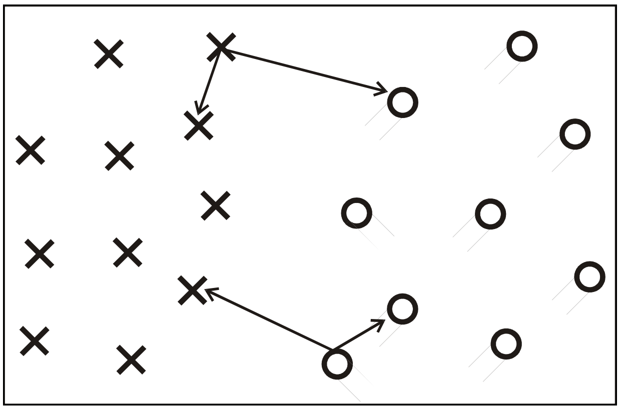


Рис. 3.1. Классификация двух образов по правилу ближайшего соседа

При использовании правила ближайшего соседа с конкретным множеством данных результирующий уровень ошибки классификации будет зависеть от случайных характеристик выборки. В частности, если для классификации $\vec{x}$ используются различные множества выборочных данных, то для ближайшего соседа вектора $\vec{x}$ будут получены различные векторы $\vec{x}_i^*$. Вместе с тем при неограниченном объеме выборки уровень ошибки по правилу ближайшего соседа никогда не будет хуже байесовского более чем в два раза.

Методы разделяющих функций

Разделяющая функция, представляемая линейной комбинацией компонент $\vec{x}$, может быть записана в виде

$$g(\vec{x}) = \vec{w}^{T} \cdot \vec{x} + w_0, \tag{3.1}$$

где $\vec{w}$ – весовой вектор; w_0 – порог.

В основу линейного классификатора для двух классов положено следующее решающее правило: если $g(\vec{x}^*) > 0$, то наблюдается класс a_1, если $g(\vec{x}^*) < 0$, то класс a_2.

Уравнение $g(\vec{x}) = 0$ определяет поверхность решений, отделяющих точки, соответствующие решению $a = a_1$, от точек, соответствующих решению $a = a_2$.

Когда функция $g(\vec{x})$ линейна, поверхность решений является гиперплоскостью. Гиперплоскость делит пространство признаков на два полупространства: область решений G_1 для a_1 и G_2 для a_2.

Если $K > 2$, то требуется несколько линейных функций и граница является кусочно-линейной. Для наглядности будем считать $K = 2$. Если на множестве объектов выполняются условия

$g(\vec{x})>0$, если $\vec{x}^0$ – реализация первого образа (a_1);

$g(\vec{x})<0$, если $\vec{x}^0$ – реализация второго образа (a_2),

то образы a_1 и a_2 называют линейно разделимыми (рис. 3.2).

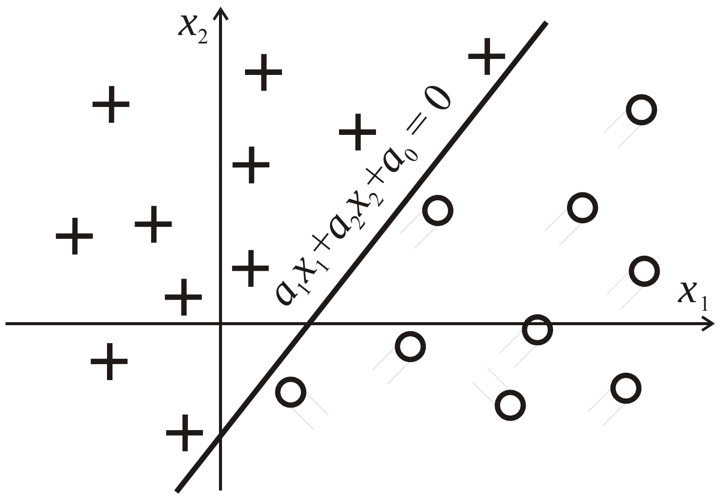


Рис. 3.2. Линейное решающее правило
для распознавания двух образов

Сложную границу раздела (для линейно неразделимых классов a_1 и a_2) можно аппроксимировать функциями вида

$$g_i(\vec{x}) = \sum_{j=0}^{n} w_{ij} \cdot x_j \,, \qquad (3.2)$$

где $x_0 \equiv 0$ – фиктивная переменная; w_{ij} – элемент матрицы весовых коэффициентов $\mathbf{W}:(p \times n+1)$.

Например, для двумерного случая $\vec{x}=(x_1, x_2)$ (рис. 3.3) граница описывается системой уравнений

$$g(x_1,x_2)= \begin{array}{ll} w_{10}+w_{11}x_1+w_{12}x_2 & \forall x_1 \in [x_1^0, x_1^1); \\ w_{20}+w_{21}x_1+w_{22}x_2 & \forall x_1 \in (x_1^1, x_1^2]; \\ w_{30}+w_{31}x_1+w_{32}x_2 & \forall x_1 \in (x_1^2, x_1^3]. \end{array}$$

Решающее правило:

если $g\left(x_1^*, x_2^*\right) > 0$, то наблюдается класс a_1, иначе – класс a_2.

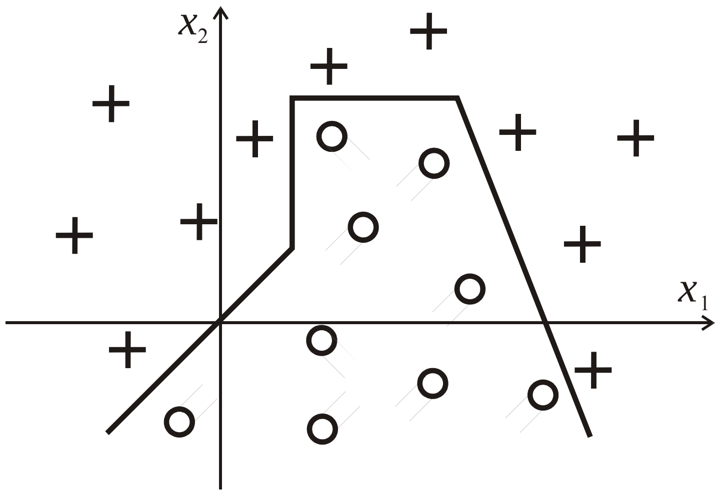


Рис. 3.3. Кусочно-линейное решающее правило для распознавания двух образов

При проведении границы между классами можно руководствоваться правилом ближайшего соседа. Если известно, что точка $\xi_1 = (x_1, y_1) \in a_1$, $\xi_2 = (x_2, y_2) \in a_2$, то $g(x, y)$ – нормаль, проведенная к отрезку, соединяющему точки ξ_1 и ξ_2, и проходящая через середину этого отрезка $M(x_0, y_0)$ с координатами

$$x_0 = \frac{x_1 + x_2}{2}, \; y_0 = \frac{y_1 + y_2}{2}.$$

В качестве справочной информации напомним необходимые сведения из векторной алгебры на плоскости.

Уравнение прямой, проходящей через две точки с координатами (x_1, y_1) и (x_2, y_2):

$$\frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1}, \text{ или } y = k \cdot x - k \cdot x_1 + y_1,$$

где $k = \dfrac{y_2 - y_1}{x_2 - x_1}$.

Уравнение перпендикуляра к прямой, проходящей через точку $M(x_0, y_0)$:

$$x - x_0 + k(y - y_0) = 0, \text{ или } y = -\frac{x - x_0}{k} + y_0.$$

Порядок выполнения работы

1. По заданным (согласно варианту) двумерным данным наблюдений $\xi_i = (x_i, y_i)$ двух классов объектов a_1 и a_2 по правилу ближайшего соседа провести границы между классами:
   - по выборочным значениям – границу $\mathrm{g1}(x, y) = 0$;
   - по выборочным средним – границу $\mathrm{g2}(x, y) = 0$.
2. Построить решающие правила $\mathrm{g1}$ и $\mathrm{g2}$.
3. Сгенерировать массивы N данных наблюдений $(\mathrm{N} = 100)$ классов a_1 и a_2 в предположении, что наблюдается двумерный случайный вектор, компоненты которого – некоррелированные нормально распределенные величины. В качестве параметров распределений классов a_1 и a_2 ($\{\hat{m}_1, \hat{\sigma}_1\}$ и $\{\hat{m}_2, \hat{\sigma}_2\}$ соответственно) взять их статистические оценки, полученные по заданным исходным данным.
4. Смоделировать процессы распознавания наблюдений по решающим правилам $\mathrm{g1}$ и $\mathrm{g2}$ и сравнить эффективности классификаторов по эмпирическим оценкам вероятностей правильных решений.
5. Оформить отчет о лабораторной работе, который должен содержать краткие теоретические сведения, алгоритмы моделирования данных и принятия решений, графические представления реализаций наблюдений и границ между классами, выводы.

Варианты заданий к лабораторной работе № 3

Номер варианта		1		2		3		4		5		6		7		8	
$\{\xi_i\}$	i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
Класс a_1	1	0	2	0	2	0	0	0	2	0	0	0	0	0	2	-1	2
	2	1	0	2	0	4	2	1	0	4	-2	2	4	2	0	0	0
	3	3	1	4	2	4	-2	3	1	8	0	4	2	4	2	0	2
Класс a_2	4	2	2	2	2	0	4	1	1	4	0	2	2	0	4	0	1
	5	3	-2	3	3	4	0	3	-2	4	2	4	0	2	2	2	-2
	6	5	1	6	3	6	3	5	1	6	0	6	2	6	3	2	1

Контрольные вопросы

1. В чем различие задач параметрического и непараметрического распознавания?
2. Назовите наиболее известные непараметрические методы распознавания. В чем их сущность?
3. Что является целью процесса «обучения с учителем»?
4. Может ли одна и та же совокупность наблюдений использоваться как обучающая и как контрольная выборки? Обоснуйте ответ.
5. Сформулируйте правило ближайшего соседа. Для решения каких задач используется это правило?
6. В чем состоит идея метода разделяющих функций?
7. Какие виды функций используются для описания границ раздела между классами?

Пример выполнения лабораторной работы

1. Исходные данные: известны результаты наблюдений двух классов объектов a_1 и a_2: $\xi_i = (x_i, y_i)$, причем $i = 1\ldots6$, $\{\xi_1, \xi_2, \xi_3\} \in a_1$, $\{\xi_4, \xi_5, \xi_6\} \in a_2$. Координаты точек ξ_i указаны в табл. 3.1.

Таблица 3.1

$\{\xi_i\}$	Класс a_1			Класс a_2		
i	1	2	3	4	5	6
x_i	-1	0	1	0	1	2
y_i	1	-1	0	0	1	-1

2. Изобразим точки ξ_i на плоскости $x0y$; для точек из разных классов будем использовать различные маркеры (рис. 3.4).

3. Проведем границу между точками $\xi_1 \in a_1$ и $\xi_4 \in a_2$ по правилу ближайшего соседа:

а) найдем координаты точки M1 – середины отрезка $[\xi_1, \xi_4]$:

$$\mathrm{M1}\ \frac{x_1+x_4}{2},\frac{y_1+y_4}{2}\ = \mathrm{M1}(-0.5, 0.5);$$

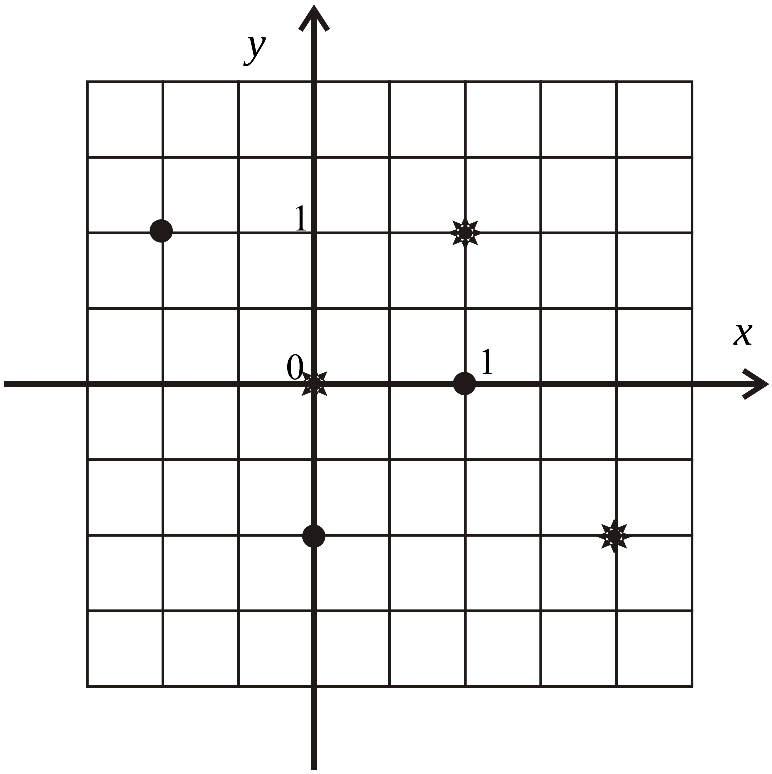


Рис. 3.4. Изображение исходных данных

б) запишем уравнение нормали к отрезку $[\xi_1, \xi_4]$, проходящей через точку M1:

$$y = -(x + 0.5) \cdot \frac{0+1}{0-1} + 0.5 = x + 1;$$

в) линия раздела между точками ξ_1 и ξ_4

$$g_1(x, y) = y - x - 1 = 0.$$

4. Проведем границу между точками $\xi_2 \in a_1$ и $\xi_4 \in a_2$. Поскольку данные точки лежат на координатной оси $0y$, то эта граница согласно правилу ближайшего соседа будет проходить перпендикулярно оси $0y$ через точку M2 – середину отрезка $[\xi_2, \xi_4]$:

$$\mathrm{M2}\ \frac{x_2 + x_4}{2}, \frac{y_2 + y_4}{2} = \mathrm{M1}(0, -0.5).$$

Следовательно, $y = -0.5$ и $g_2(x, y) = y + 0{,}5 = 0$.

5. Точка пересечения линий $g_1(x, y) = 0$ и $g_2(x, y) = 0$ – это точка излома S_1 кусочно-линейной границы $g1(x, y) = 0$ между классами a_1 и a_2; координаты этой точки находим из решения системы уравнений:

$$y = x + 1;$$
$$y = -0.5.$$

Тогда $S_1 = S_1(-1.5, -0.5)$.

6. Аналогичным образом находим уравнения линий, разделяющих остальные пары точек, и координаты точек излома $S_k(x_k, y_k)$ границы $g1(x, y) = 0$. Результаты расчетов представлены в табл. 3.2.

Вид границы $g1$ между классами показан на рис. 3.5.

Таблица 3.2

k	0	1	2	3	4	5	
$g1(x, y)$	$x = 0$	$y = x + 1$	$y = -0.5$	$x = 0.5$	$y = 0.5$	$y = x - 2$	$x = 1$
$S_k(x_k, y_k)$	(0, 1)	(–1.5,–0.5)	(0.5,–0.5)	(0.5, 0.5)	(2.5, 0.5)	(1, –1)	

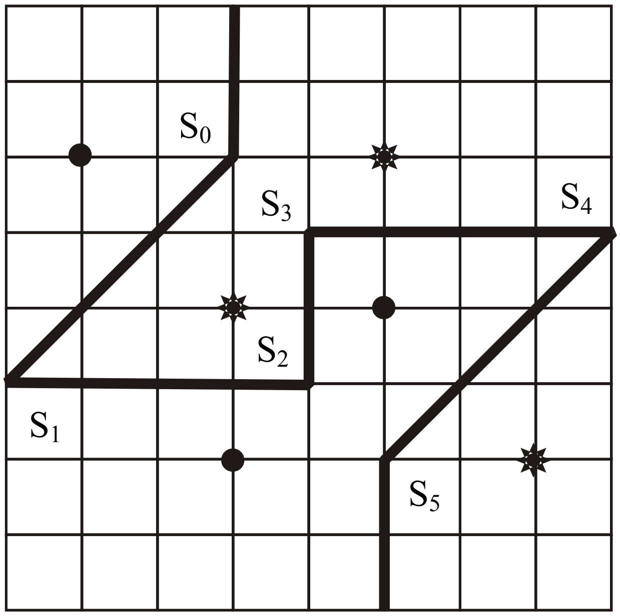


Рис. 3.5. Граница между классами $g1$

7. Составим алгоритм классификации объектов a_1 и a_2 с использованием кусочно-линейной границы $g1(x, y) = 0$.

Обозначим γ_1 – принятие решения о том, что наблюдается класс a_1, γ_2 – принятие решения о том, что наблюдается класс a_2. Линии $x = x_k$ (проходящие через точки S_k параллельно оси $0y$) делят плоскость $x0y$ на области G_k принятия решения по следующим правилам:

{если $x < -1.5$, тогда γ_1,
иначе если $x \in [-1.5,\ 0]$, то
если $-0.5 \le y \le x + 1$, тогда γ_2, иначе γ_1,
иначе если $x \in (0,\ 0.5]$, то
если $y \ge -0.5$, тогда γ_2, иначе γ_1, (3.3)
иначе если $x \in (0.5,\ 1]$, то
если $y \ge 0.5$, тогда γ_2, иначе γ_1,
иначе если $x \in (1,\ 2.5]$, то
если $x - 2 \le y \le 0.5$, тогда γ_1, иначе γ_2,
иначе γ_2}.

8. Поскольку данные наблюдений – случайные величины $\vec{\xi}_i$, можно провести границу между классами после статистической обработки данных – нахождения средних выборочных значений:

$$\vec{m}_\xi \mid a_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \vec{\xi}_j \Big| \vec{\xi}_j \in a_i ,$$

где $\vec{m}_\xi \mid a_i$ – вектор средних значений координат $(x_j,\ y_j)$ двумерных данных $\vec{\xi}_j$, принадлежащих классу a_i, ($i = 1,\ 2$ – номер класса), n_i – объем выборки данных класса a_i.

Определим компоненты векторов $\vec{m}_1 = (\mathrm{m1}_0, \mathrm{m1}_1)$ и $\vec{m}_2 = (\mathrm{m2}_0, \mathrm{m2}_1)$ – статистических оценок математических ожиданий (МО) классов a_1 и a_2:

$$\{\mathrm{m1}_0 := \frac{1}{3}\cdot\sum_{\mathrm{i}=1}^{3}\mathrm{x_i},\ \mathrm{m1}_1 := \frac{1}{3}\cdot\sum_{\mathrm{i}=1}^{3}\mathrm{y_i}\}\ \text{ и }\ \{\mathrm{m2}_0 := \frac{1}{3}\cdot\sum_{\mathrm{i}=4}^{6}\mathrm{x_i},\ \mathrm{m2}_1 := \frac{1}{3}\cdot\sum_{\mathrm{i}=4}^{6}\mathrm{y_i}\};$$

получим $\vec{m}_1 = (0,\ 0)$, $\vec{m}_2 = (1,\ 0)$.

9. Найдем уравнение границы g2 между классами a_1 и a_2 после усреднения данных наблюдений. Эта граница проходит через точку с координатами $\frac{0+1}{2}, \frac{0+0}{2} = (0.5,\ 0)$ и параллельна оси $0y$:

$$\mathrm{g2}(x, y) = x - 0.5 = 0.$$

Соответствующее правило принятия решения будет таким:

{если $x \le 0.5$, тогда γ_1, иначе γ_2} (3.4)

Расположение данных, их средних значений ($\vec{m}_1$ и $\vec{m}_2$) и граница между классами показаны на рис. 3.6.

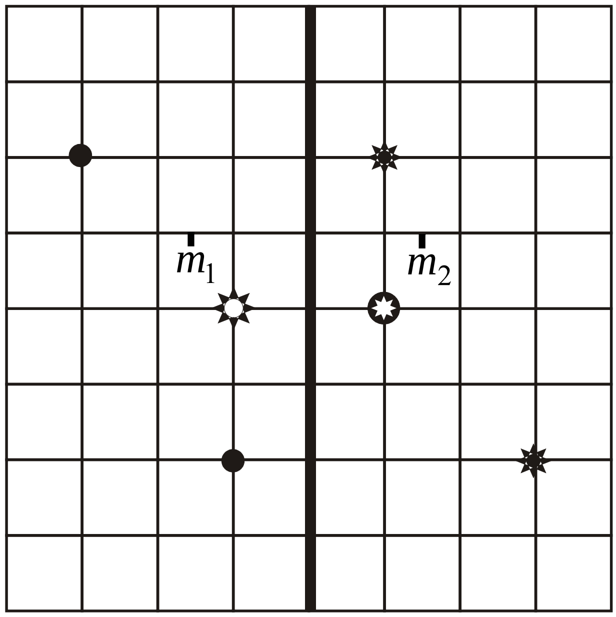


Рис. 3.6. Граница между классами $g2$

10. Для экспериментальной проверки качества работы классификаторов по правилам (3.3) и (3.4) смоделируем результаты наблюдений – массивы $(\{x1_i\};\ \{y1_i\})$ и $(\{x2_i\};\ \{y2_i\})$, соответствующие классам a_1 и a_2.

Подробно рассмотрим только случай класса a_1.

Считаем, что $x1$ и $y1$ – некоррелированные компоненты двумерной случайной величины, подчиняющейся нормальному закону распределения с МО $\vec{m}1 = (m1_0, m1_1)$ и среднеквадратическим отклонением (СКО) $\vec{\sigma}1 = (\sigma 1_0, \sigma 1_1)$. В качестве значений параметров распределения примем их статистические оценки:

МО: $m1_0 = 0$; $m1_1 = 0$ (см. п. 8);

дисперсия:

$$Dl_0 := \frac{1}{2} \cdot \sum_{i=1}^{3} (x_i - m1_0)^2 ; \quad Dl_1 := \frac{1}{2} \cdot \sum_{i=1}^{3} (y_i - m1_1)^2 ;$$

СКО: $\sigma 1_0 := \sqrt{Dl_0}$; $\sigma 1_1 := \sqrt{Dl_1}$;

$\sigma 1_0 = 1$; $\sigma 1_1 = 1$.

Определим пользовательскую функцию, осуществляющую алгоритм генерации массива реализаций нормально распределенной случайной величины:

$$n := 48 \quad k := 1..n$$

$$\mathrm{Norm}(z,m,\sigma) := \sqrt{\frac{12}{n}} \cdot \sigma \cdot \left(\sum_k \mathrm{rnd}(1) - \frac{n}{2} \right) + m.$$

Формальными аргументами этой функции являются номера элементов z массива реализаций и параметры нормального распределения МО m и СКО σ моделируемой случайной величины.

Получим 100 данных наблюдений класса a_1:

$$N := 100 \quad i := 0..N-1$$

$$x1_i := \mathrm{Norm}(i, m1_0, \sigma1_0) \quad y1_i := \mathrm{Norm}(i, m1_1, \sigma1_1).$$

Аналогичным образом сгенерируем 100 данных наблюдений класса a_2: $x2_i$ и $y2_i$.

11. Построим графическое изображение данных (рис. 3.7).

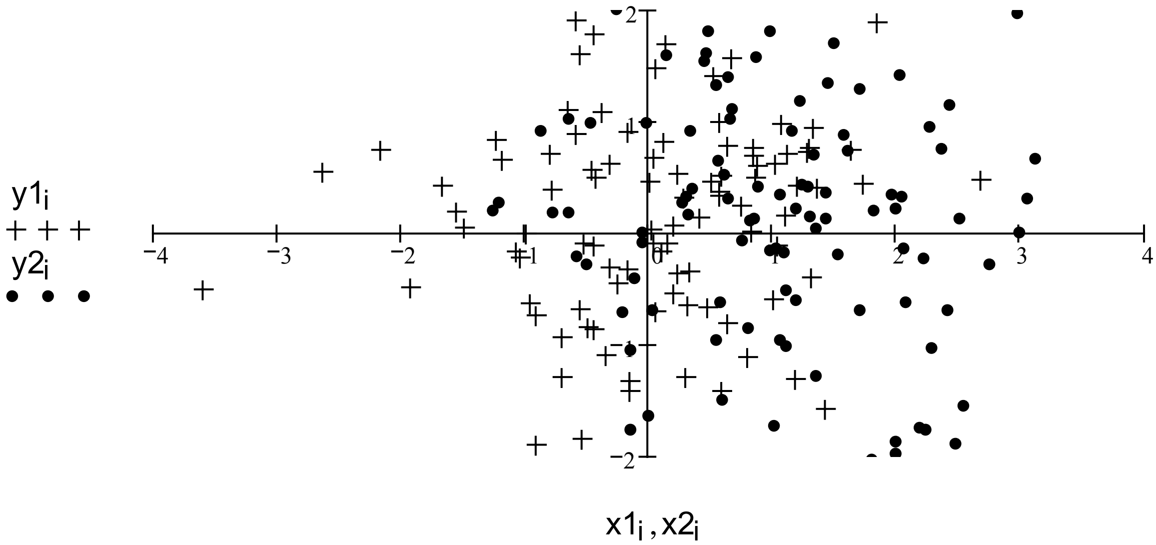


Рис. 3.7. Реализации наблюдений

12. Выполним распознавание контрольной выборки $(\{x_i\};\ \{y_i\})$ по решающему правилу (3.3).

Пусть контрольная выборка принадлежит классу a_1, тогда

$$x_i := x1_i \quad y_i := y1_i.$$

Формализуем описание процедуры принятия решения (3.3):

$$a_i := \begin{array}{|ll} 1 & \text{if } x_i < -1.5 \\ a_i \leftarrow \mathrm{if}[-0.5 \le y_i \le (x_i + 1), 2, 1] & \text{if } -1.5 \le x_i \le 0 \\ a_i \leftarrow \mathrm{if}(y_i \ge -0.5, 2, 1) & \text{if } 0 < x_i \le 0.5 \\ a_i \leftarrow \mathrm{if}(y_i \ge 0.5, 2, 1) & \text{if } 0.5 < x_i \le 1 \\ a_i \leftarrow \mathrm{if}[(x_i - 2) \le y_i \le 0.5, 1, 2] & \text{if } 1 < x_i \le 2.5 \\ 2 & \text{otherwise} \end{array}.$$

В результате получим массив a_i, элементы которого равны либо 1 (если принято решение γ_1), либо 2 (если принято решение γ_2). Поскольку распознавался класс a_1, то γ_1 – правильное решение, γ_2 – ошибочное.

Определим эмпирическую вероятность правильного распознавания класса a_1 как отношение количества правильных решений к объему испытаний N:

$$P11 := \frac{1}{N} \cdot \sum_i \mathrm{if}(a_i = 1, 1, 0).$$

Тогда эмпирическая вероятность ошибочного распознавания

$$P21 := 1 - P11.$$

Получим: $P11 = 0.6$; $P21 = 0.4$.

13. Проведем распознавание контрольной выборки $(\{x_i\};\ \{y_i\})$ по решающему правилу (3.4):

$$a_i := \mathrm{if}(x_i < 0.5, 1, 2).$$

Оценим вероятности $P11$ и $P21$.

Получим: $P11 = 0.76$; $P21 = 0.24$.

14. Выполним пп. 12, 13 для случая, когда контрольная выборка принадлежит классу a_2.

15. Сравним эмпирические оценки эффективности классификации данных по решающим правилам (3.3) и (3.4).

В выводах по лабораторной работе следует отметить, как влияют на результаты распознавания следующие факторы: предположение о законе распределения данных, объемы обучающих и контрольной выборок, способ формирования границы раздела.

Лабораторная работа № 4

МЕТОДЫ ГРУППИРОВКИ ДАННЫХ

Цель работы:

– изучить основные принципы «обучения без учителя» и методики группировки данных в условиях полной апостериорной неопределенности;

– получить навыки иерархической группировки данных с применением различных мер внутриклассового расстояния.

Теоретические сведения

Наиболее проблематична задача построения решающего правила в условиях полной апостериорной неопределенности: когда имеется множество выборочных значений (представленных в виде файлов данных наблюдений или изображений) без указания их классификации, т.е. заранее не известно ни количество объектов, ни что это за объекты. Поэтому первый шаг процесса *обучения без учителя* – разделить данные в подгруппы (*кластеры*); при этом в одну группу объединяются данные с похожими признаками. Сразу возникают два вопроса: как измерять сходство между наблюдениями (отсчетами) и как оценивать разделение выборочного множества на группы. Наиболее очевидной мерой подобия (или различия) между двумя отсчетами является расстояние между ними: расстояние между отсчетами в одной группе (одном классе) будет существенно меньше, чем расстояние между отсчетами из разных групп.

Предположим, что два отсчета $\vec{x}$ и $\vec{x}'$ принадлежат одной группе, если *евклидово расстояние* между ними

$$r=\sqrt{(\vec{x}-\vec{x}')^{\mathrm{T}}(\vec{x}-\vec{x}')} = \left\|\vec{x}-\vec{x}'\right\| \quad (4.1)$$

меньше, чем пороговое расстояние d_0. Значение d_0 должно быть больше, чем типичные внутригрупповые расстояния, и меньше, чем типичные межгрупповые расстояния.

В случае двумерного признакового пространства $\vec{x}^{\mathrm{T}}=(x_1,x_2)$, $\vec{x}'^{\mathrm{T}}=(x_1',x_2')$ формула (4.1) имеет вид

$$r = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2}\,.$$

Качество группировки любой части данных определяется с помощью *функций критерия*. Самая простая и наиболее используемая функция критерия – сумма квадратов ошибок:

$$J_e = \sum_{i=1}^{k} \sum_{\vec{x} \in X_i} \left\| \vec{x} - \vec{m}_i \right\|^2, \tag{4.2}$$

где $\vec{m}_i$ – средний вектор i-й группы, состоящей из n_i отсчетов:

$$\vec{m}_i = \frac{1}{n_i} \sum_{\vec{x} \in X_i} \vec{x}\,. \tag{4.3}$$

Интерпретация (4.2): для данной группы X_i средний вектор $\vec{m}_i$ лучше всего представляет отсчеты в X_i, так как он минимизирует сумму квадратов длин векторов «ошибок» – отклонений $\vec{x}$ от $\vec{m}_i$. Таким образом, критерий J_e измеряет общую квадратичную ошибку, вносимую при представлении n отсчетов $\vec{x}_1, \ldots, \vec{x}_n$ центрами k групп $\vec{m}_1, \ldots, \vec{m}_k$. Оптимальным разделением считается то, которое минимизирует J_e. Группировки такого рода называют разделением с минимальной дисперсией. Обычно J_e – подходящий критерий, если отсчеты образуют облака, которые хорошо отделены друг от друга.

Функции критерия можно получить из матриц рассеяния, используемых в дискриминантном анализе:

– матрица рассеяния для i-й группы:

$$\mathbf{S}_i = \sum_{\vec{x} \in X_i} (\vec{x} - \vec{m}_i)^{\mathrm{T}} (\vec{x} - \vec{m}_i);$$

– матрица рассеяния внутри группы:

$$\mathbf{S}_W = \sum_{i=1}^{k} \mathbf{S}_i\,;$$

– матрица рассеяния между группами:

$$\mathbf{S}_B = \sum_{i=1}^{k} n_i (\vec{m}_i - \vec{m})^{\mathrm{T}} (\vec{m}_i - \vec{m}),$$

где $\vec{m} = \frac{1}{n} \sum_{i=1}^{k} n_i \vec{m}_i$ – общий средний вектор;

– общая матрица рассеяния:

$$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B\,.$$

Общая матрица рассеяния зависит только от общего объема выборки. Внутригрупповые и межгрупповые матрицы рассеяния зависят от того, как множество отсчетов разделено на группы. Чтобы оценить степень внутри- и межгруппового рассеяния, вводят скалярную меру матриц рассеяния: след и определитель.

Когда найдена функция критерия, группировка становится корректно поставленной задачей дискретной (k) оптимизации: найти такие разделения выборочного множества, которые приводят к экстремуму функции критерия. Наиболее часто для поиска оптимального разделения используется итеративная оптимизация, основная идея которой заключается в нахождении некоторого начального разделения и «передвижении» отсчетов из одной группы в другую, если это передвижение улучшает значение функции критерия. Такой подход гарантирует, в общем случае, локальный, но не глобальный экстремум; различные начальные точки могут привести к различным решениям, при этом не известно, было ли найдено лучшее решение.

Рассмотрим последовательность разделений n отсчетов на k групп. Первое – это разделение на n групп, причем каждая группа содержит по одному отсчету. Затем – разделение на $(n-1)$ группу, на $(n-2)$ группы и т.д. Если два отсчета образуют одну группу на уровне k и остаются вместе на более высоких уровнях, то такую последовательность называют ***иерархической группировкой***. *Агломеративные* (объединяющие) процедуры начинают с n одиночных групп и образуют последовательность постепенно объединяемых групп. *Делимые* процедуры начинают с одной группы, содержащей все n отсчетов, и образуют последовательность постепенно разделяемых групп.

Основные шаги ***базовой агломеративной группировки***:

1) пусть текущее число групп $\hat{k} = n$ и $X_i = \{\bar{x}_i\}$, $i = 1,\ldots, n$;

ЦИКЛ: 2) если $\hat{k} \leq k$ (заданное число групп), то СТОП;

3) найти ближайшую пару групп (X_i и X_j);

4) объединить X_i и X_j, уничтожить X_j и уменьшить $\hat{k}$ на 1;

5) перейти к шагу ЦИКЛ.

В качестве мер расстояния между двумя группами используют следующие критерии:

$$d_{\min}(X_i, X_j) = \min_{\substack{\vec{x} \in X_i, \\ \vec{x}' \in X_j}} \|\vec{x} - \vec{x}'\|; \qquad (4.4)$$

$$d_{\max}(X_i, X_j) = \max_{\substack{\vec{x} \in X_i, \\ \vec{x}' \in X_j}} \|\vec{x} - \vec{x}'\|; \qquad (4.5)$$

$$d_{\text{mean}}(X_i, X_j) = \|\vec{m}_i - \vec{m}_j\|. \qquad (4.6)$$

Все эти меры напоминают минимальную дисперсию (4.2) и обычно дают одинаковые результаты, если группы компактны и хорошо разделены. Однако, если группы близки друг к другу или их форма не гиперсферическая, результаты группировок с использованием различных критериев (4.4) – (4.6) могут быть разными.

Алгоритм «ближайший сосед»

Рассмотрим случай, когда используется $d_{\min}$. Предположим, что точки данных рассматриваются как вершины графа, причем ребра графа образуют путь между вершинами в одном подмножестве X. Когда для измерения расстояния между подмножествами используется $d_{\min}$, ближайшие соседи определяют ближайшие подмножества. Слияние X_i и X_j соответствует добавлению ребра между двумя ближайшими вершинами в X_i и X_j. Поскольку ребра, соединяющие точки подмножества, всегда проходят между различными группами, то результирующий граф никогда не будет иметь замкнутых контуров или цепей. Пользуясь терминологией теории графов, можно сказать, что эта процедура генерирует дерево. Если ее продолжать до тех пор, пока все точки подмножеств не будут соединены, то в результате получим покрывающее дерево (остов) – дерево с путем от любой вершины к любой другой вершине в группе. При этом сумма длин ребер результирующего дерева не будет превышать сумм длин ребер для любого другого покрывающего дерева для данного множества выборочных данных. Таким образом, используя $d_{\min}$ в качестве меры расстояния, агломеративная процедура группировки превращается в алгоритм для генерирования минимального покрывающего дерева.

Минимальное покрывающее дерево можно получить, добавляя самое короткое ребро между двумя другими ребрами (двумя ближайшими парами точек).

Если некоторые точки расположены так, что между исходными группами создается некоторый мост, то это приводит к «цепному эффекту» – объединению данных в одну большую продолговатую группу и одну или несколько маленьких компактных групп. Такой эффект наблюдается, когда результаты очень чувствительны к шуму или к небольшим изменениям в положении точек данных. Однако та же тенденция формирования цепей может считаться преимуществом, если группы сами по себе вытянуты или имеют вытянутые отростки.

Алгоритм «дальний сосед»

Когда для измерения расстояния между группами используется $d_{\max}$, возникновение вытянутых групп является нежелательным. Применение процедуры можно рассматривать как получение графа, в котором ребра соединяют все вершины в группу. Пользуясь терминологией теории графов, можно сказать, что каждая группа образует полный подграф. Расстояние между двумя группами определяется наиболее удаленными вершинами в этих двух группах. Когда две ближайшие группы объединяются, граф изменяется добавлением ребер между каждой парой вершин в двух объединяемых группах. Расстояние между двумя группами определяется наиболее удаленными вершинами в этих двух группах. Если *диаметр группы* определяется как наибольшее расстояние между точками в группе, то расстояние между двумя группами – просто диаметр их объединения. Если *диаметр разделения* определяется как наибольший диаметр для группы в разделении, то каждая итерация увеличивает диаметр разделения минимально. Это является преимуществом в том случае, когда истинные группы компактны и примерно одинаковы по размерам. Однако в других случаях, как, например, в случае вытянутых групп, результирующая группировка бессмысленна. Это еще один пример наложения структуры на данные вместо нахождения их структуры.

Порядок выполнения работы

1. Для заданных (согласно варианту) двумерных данных наблюдений $\xi_i = (x_i, y_i)$, $i = 1, 2, \ldots, 10$, в условиях полной априорной неопределенности установить меры подобия – вычислить расстояния от

каждой точки из заданного множества ξ_i до всех других точек ξ_j, $i \neq j$.

2. Построить полигон распределения расстояний d_{ij} и найти центры рассеяния.

3. Установить критерии объединения данных $d_{\min}$ и $d_{\max}$.

4. Выполнить группировки данных по алгоритму ближайшего соседа и по алгоритму дальнего соседа.

5. Определить статистические оценки характеристик разброса внутри классов: минимальное, максимальное и среднее расстояния между парами точек, составляющих одну группу (один класс).

6. Выполнить сравнительный анализ результатов группировок по алгоритмам ближайшего соседа и дальнего соседа.

7. Оформить отчет о лабораторной работе, который должен содержать алгоритмы группировки, результаты расчетов, графические представления исходных данных и полученных групп, выводы.

Варианты заданий к лабораторной работе № 4

Номер варианта	i	1	2	3	4	5	6	7	8	9	10
1	x_i	0	1	2	3	4	4	6	6	8	8
	y_i	3	0	3	1	2	5	2	5	2	6
2	x_i	0	1	2	2	3	3	4	5	6	6
	y_i	2	3	0	3	1	4	5	2	2	3
3	x_i	0	1	2	2	4	4	4	5	6	6
	y_i	2	4	0	4	1	4	5	3	3	4
4	x_i	1	2	3	3	4	4	5	5	7	8
	y_i	3	3	0	4	1	6	3	7	3	4
5	x_i	0	1	2	3	3	5	5	7	7	8
	y_i	1	3	-1	-1	2	0	3	0	3	-1
6	x_i	0	2	2	4	4	6	6	8	8	9
	y_i	2	-1	4	-1	3	0	3	-1	4	1
7	x_i	0	1	1	2	2	3	3	4	5	6
	y_i	0	-1	3	0	2	0	3	1	3	1
8	x_i	1	2	3	4	5	5	6	6	7	8
	y_i	1	4	0	5	1	4	0	1	4	3

Контрольные вопросы

1. В чем различие процессов «обучения с учителем» и «обучения без учителя»?

2. Как измерять сходство между данными наблюдений?

3. Дайте определения различных мер внутриклассового расстояния и мер расстояния между классами.

4. Какие методики группировки данных Вам известны?

5. Как оценить качество группировки любой части данных? Приведите примеры функций критериев группировки.

6. Проанализируйте недостатки и преимущества алгоритмов ближайшего соседа и дальнего соседа. Как выбрать наиболее подходящий алгоритм для группировки данных?

Пример выполнения лабораторной работы

1. Исходные данные: имеется множество результатов наблюдений $\xi_i = (x_i, y_i)$, $i = 1 \ldots 10$, без указания их классификации. Количество наблюдаемых классов не известно.

Координаты точек ξ_i указаны в табл. 4.1.

Таблица 4.1

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	0	0	1	1	2	2	3	3	4	4
y_i	0	3	1	3	1	4	2	4	1	5

Условимся для обозначения конкретной точки ξ_i из множества данных использовать ее порядковый номер i согласно табл. 4.1.

2. Нанесем точки на координатную плоскость $x0y$, как показано на рис. 4.1.

3. Определим расстояния d_{ij} между всеми возможными парами точек $i = 1 \ldots 10, j = 1 \ldots 10$:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}\ .$$

Результаты расчетов представлены в табл. 4.2.

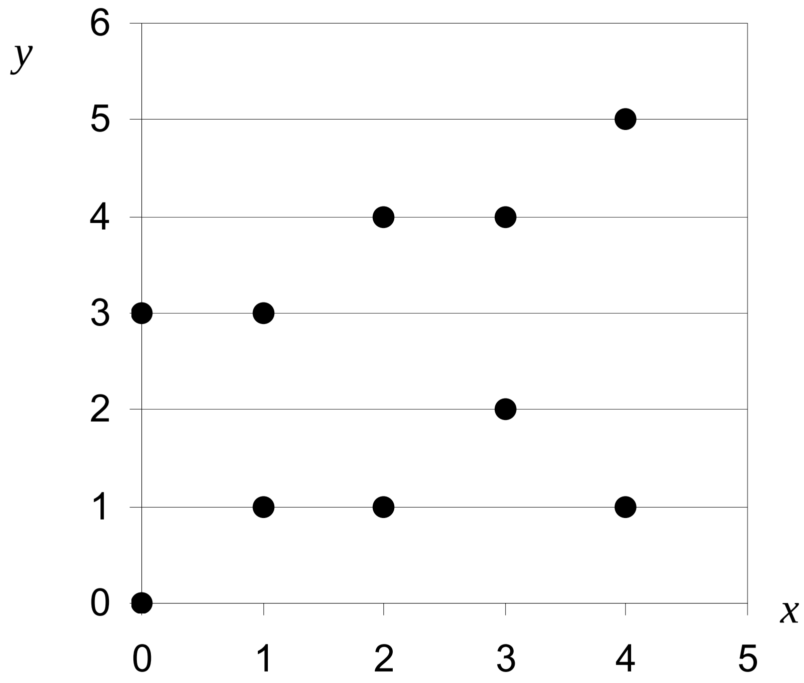


Рис. 4.1. Изображение исходных данных

4. Полученные значения расстояний d_{ij} можно рассматривать как N ($N = 45$ – объем выборки) независимых значений одной случайной величины d.

Упорядочим расстояния d_{ij} по возрастанию. Результаты сортировки представим в виде табл. 4.3, где k – количество пар точек, расстояние между которыми равно d.

Чтобы упростить статистическую обработку данных, выберем в качестве середин *разрядов* ряда распределения d повторяющиеся результаты расчетов расстояний d_i, ($i = 1\ldots13$ – номер разряда) (см. табл. 4.3).

Для более наглядного представления о величинах расстояний между точками выборочного множества построим полигон распределения расстояний. **Полигон** (рис. 4.2) – это ломаная линия, соединяющая точки с абсциссами d_i (середины разрядов) и ординатами r_i (относительными частотами). *Относительной частотой* (иначе *эмпирической вероятностью*) называется число наблюдений в i-м разряде k_i, отнесенное к объему выборки N:

$$r_i = k_i / N\,.$$

Таблица 4.2

d_{ij}	2	3	4	5	6	7	8	9	10
1	3.000	1.414	3.162	2.236	4.472	3.606	5.000	4.123	6.403
2		2.236	1.000	2.828	2.236	3.162	3.162	4.472	4.472
3			2.000	1.000	3.162	2.236	3.606	3.000	5.000
4				2.236	1.414	2.236	2.236	3.606	3.606
5					3.000	1.414	3.162	2.000	4.472
6						2.236	1.000	3.606	2.236
7							2.000	1.414	3.162
8								3.162	1.414
9									4.000

Таблица 4.3

d_i	1.000	1.414	2.000	2.236	2.828	3.000	3.162
k_i	3	5	3	9	1	3	7

d_i	3.606	4.000	4.123	4.472	5.000	6.403
k_i	5	1	1	4	2	1

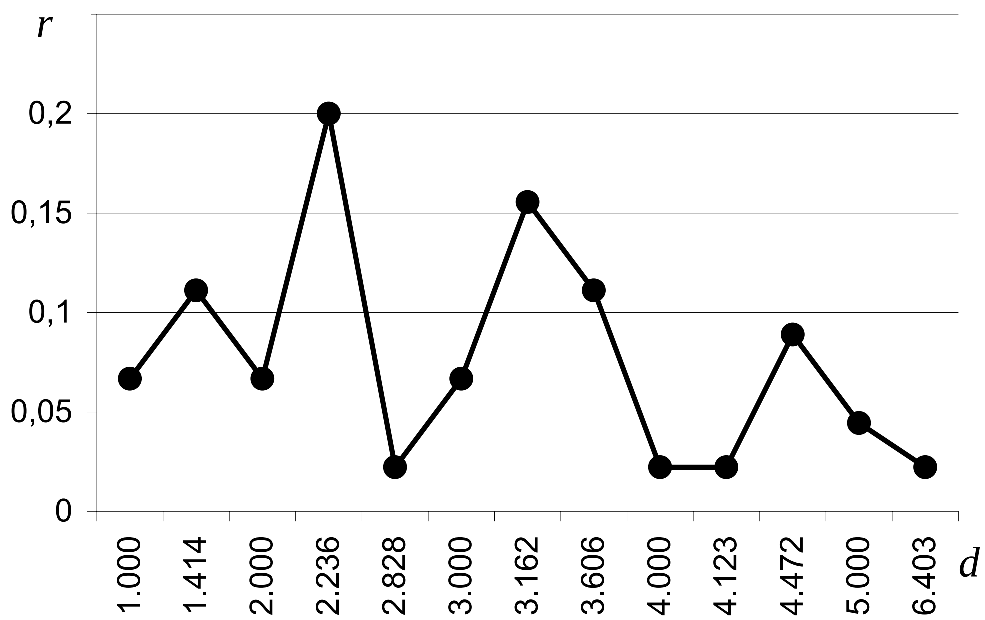


Рис. 4.2. Полигон распределения расстояний между парами точек

5. Проведем предварительный анализ меры рассеяния d_{ij} выборочного множества. Наиболее удалены друг от друга точки (1) и (10), предполагаем, что они относятся к различным классам a_1 и a_2.

Пусть для определенности $(1) \in a_1$, $(10) \in a_2$.

В качестве начальной точки алгоритмов иерархической группировки возьмем (1).

6. Выполним группировку данных по алгоритму ближайшего соседа. Последовательность шагов алгоритма:

- найдем точку, ближайшую к точке (1). Для этого в строке № 1 табл. 4.2 найдем минимальный элемент d_{1j} ; это $d_{13} = 1.414$;
- объединяем точки (1) и (3) в одну группу (класс a_1) и переходим к следующей точке группы – (3);
- находим минимальный элемент d_{3j} в строке № 3; $\min\{d_{3j}\} = d_{35} = 1.000$; считаем, что $(5) \in a_1$, и переходим на строку № 5;
- в строке № 5 $\min\{d_{5j}\} = d_{57} = 1.414$, относим точку (7) к классу a_1 и переходим на строку № 7;
- в строке № 7 $\min\{d_{7j}\} = d_{79} = 1.414$, $(9) \in a_1$;
- в строке № 9 находится единственный элемент – $d_{9,10}$, и мы не можем отнести точку (10) к группе точек $\{1, 3, 5, 7, 9\}$, поскольку изначально предполагали, что точки (1) и (10) относятся к разным классам. Следовательно, обрываем цепочку последовательных объединений точек в одну группу.
- переходим к одной из точек, не вошедших в первую группу (класс a_1), и повторяем шаги алгоритма.

Таким образом, в результате работы алгоритма ближайшего соседа получаем следующее разделение данных:

класс a_1: $\{1, 3, 5, 7, 9\}$, класс a_2: $\{2, 4, 6, 8, 10\}$.

7. Определим статистические характеристики рассеяния внутри каждого класса:

- класс a_1: минимальное расстояние $d_{35} = 1.00$, максимальное – $d_{19} = 4.12$, среднее – $d_{a1} = 2.24$;
- класс a_2: минимальное расстояние $d_{24} = d_{68} = 1.00$, максимальное – $d_{2,10} = 4.47$, среднее – $d_{a2} = 2.28$.

8. Для визуализации результатов группировки соединим линиями пары точек, принадлежащих одному и тому же классу (рис. 4.3, а).

9. Выполним группировку данных по алгоритму дальнего соседа.

Считаем, что количество классов объектов не менее двух $(K \geq 2)$ и граничные точки (1) и (10) принадлежат разным классам: $(1) \in a_1$, $(10) \in a_2$. Для работы алгоритма необходимо оценить диаметр группы – наибольшее расстояние между точками одной группы. На рис. 4.2 визуально выделяются четыре моды рассеяния точек:

– мода $d^1 = 1.414$ соответствует в основном несвязанным между собой парам точек: (1–3), (4–6), (8–10), т.е. указанные ребра не имеют общих вершин;

– мода $d^2 = 2.236$ – это расстояние между связанными парами: {(4–5), (4–7), (4–8)}, {(6–2), (6–7), (6–10)}, {(7–3), (7–4), (7–6)}; точки (4), (6) и (7) являются центрами рассеяния с мерой d^2.

Сравним расстояния от центров рассеяния (4), (6), (7) до граничных точек (1) и (10) (см. табл. 4.1):

$$d_{14} < d_{17} < d_{16}\ (3.16 < 3.61 < 4.47);$$

$$d_{10,6} < d_{10,7} < d_{10,4}\ (2.24 < 3.16 < 3.61);$$

в качестве критерия *разброса внутри класса* $d_{\max}$ примем d_{14};

– мода $d^3 = 3.126$ – расстояние между точками {(8–2), (8–5), (8–9)}; центр рассеяния – точка (8). Ближайшая к (8) граничная точка – точка (10) – находится на расстоянии $d_{8,10} = 1.414 < d_{\max} = 3.162$, поэтому точки (10) и (8) относим к одному классу (a_2);

– мода $d^4 = 4.472$ характеризует разброс между точками из разных классов; это означает, что пары (2)–(9), (2)–(10) и (5)–(10) образуют группы {2, 5} и {9, 10}.

Сгруппируем данные по такому правилу: расстояние от граничной точки до любой другой точки группы не должно превышать $d_{\max} = 3.162$.

Упорядочим расстояния d_{1i}, $d_{10,i}$ $(i = 1\ldots10)$ (табл. 4.4).

Как видно из таблицы, по критерию $d_{\max}$ выделяются две группы: $\{1, 2, 3, 4, 5\} \in a_1$ и $\{6, 7, 8, 10\} \in a_2$ и отдельная точка (9). Ближайшая к точке (9) – точка (7) и $d_{9,10} = 4.00 < d_{9,1} = 4.123$, поэтому отнесем точку (9) к классу a_2.

Таблица 4.4

	Точка (10)								
i	8	6	7	4	9	2	5	3	1
d_{ij}	1.414	2.236	3.162	3.606	4.000	4.472	4.472	5.000	6.403

	Точка (1)								
i	3	5	2	4	7	9	6	8	10
d_{ij}	1.414	2.236	3.000	3.162	3.606	4.123	4.472	5.000	6.403

Таким образом, по алгоритму дальнего соседа получаем:

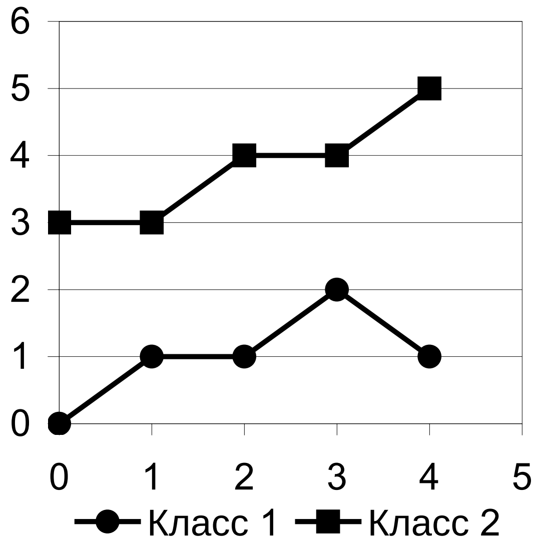
класс a_1: $\{1, 2, 3, 4, 5\}$, класс a_2: $\{6, 7, 8, 9, 10\}$.

10. Оценим характеристики рассеяния внутри каждого класса:

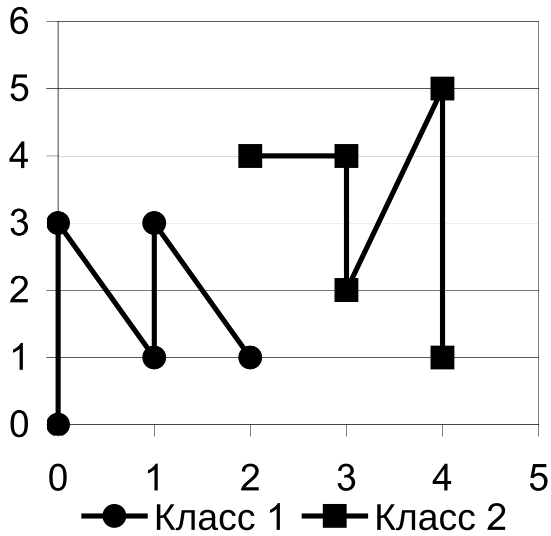
– класс a_1: минимальное расстояние $d_{35} = d_{24} = 1.00$, максимальное – $d_{14} = 3.16$, среднее – $d_{a1} = 2.11$;

– класс a_2: минимальное расстояние $d_{68} = 1.00$, максимальное – $d_{9,10} = 4.00$, среднее – $d_{a2} = 2.42$.

11. Результаты группировки точек показаны на рис. 4.3, б.



а



б

Рис. 4.3. Изображение результатов группировки данных:
а – алгоритм ближайшего соседа; б – алгоритм дальнего соседа

В выводах по лабораторной работе необходимо проанализировать влияние выбора алгоритма на результаты группировки данных.

Библиографический список

Бабаков М.Ф. Математические модели электронных аппаратов и систем: учеб. пособие / М.Ф. Бабаков, А.В. Попов, М.И. Луханин. – Х.: Нац. аэрокосм. ун-т "Харьк. авиац. ин-т", 2003. – 109 с.

Бабаков М.Ф. Методы машинного моделирования в проектировании электронной аппаратуры: учеб. пособие / М.Ф. Бабаков, А.В. Попов. – Х.: Нац. аэрокосм. ун-т "Харьк. авиац. ин-т", 2002. – 89 с.

Гмурман В.Е. Теория вероятностей и математическая статистика / В.Е. Гмурман. – М.: Высш. шк., 2000. – 479 с.

Дуда Р. Распознавание образов и анализ сцен: пер. с англ. / Р. Дуда, П. Харт; под ред. В.Л. Стефанюка. – М.: Мир, 1976. – 511 с.

Пересада В.П. Автоматическое распознавание образов / В.П. Пересада. – Л.: Энергия, 1970. – 90 с.

Фомин Я.А. Статистическая теория распознавания образов / Я.А. Фомин, Г.Р. Тарловский. – М.: Радио и связь, 1986. – 264 с.

Фукунага К. Введение в статистическую теорию распознавания образов: пер. с англ. / К. Фукунага; под ред. А.А. Дорофеюка. – М.: Наука, 1979. – 367 с.

Содержание

Васильева Ирина Карловна
Ельцов Павел Евгеньевич

МЕТОДЫ РАСПОЗНАВАНИЯ ОБРАЗОВ

Редактор Е.Ф. Сережкина

Св. план, 2008
Подписано в печать
Формат 60×84 1/16. Бум. офс. № 2. Офс. печ.
Усл. печ. л. 3 . Уч.-изд. л. . Т. 50 экз. Заказ . Цена свободная

Национальный аэрокосмический университет им. Н.Е. Жуковского
"Харьковский авиационный институт"
61070, Харьков-70, ул. Чкалова, 17
http://www.khai.edu
Издательский центр “ХАИ”
61070, Харьков-70, ул. Чкалова, 17
izdat@khai.edu