Использование ковариационной матрицы для вычисления весовых коэффициентов классификаторов в задаче многоклассового распознавания

Using covariation matrix for calculating classifiers' weight coefficients in multi class recognition task

Субботенко / Subbotenko A.
Александр Владимирович
(subbiki@yandex.ru)
кандидат технических наук.
Академия ФСО России,
сотрудник.
г. Орел

Петрухин / Petrukhin G.
Геннадий Дмитриевич
(subbiki@yandex.ru)
доктор технических наук, профессор.
В/ч 33965, сотрудник.
г. Москва

Кирюхин / Kiryukhin D.
Дмитрий Александрович
(mestnii81@rambler.ru)
кандидат технических наук.
Академия ФСО России,
сотрудник.
г. Орел

Ключевые слова: распознавание – recognition; ковариационная матрица – covariation matrix; «след» матрицы – matrix 'trace'; весовой коэффициент – weight coefficient.

В статье рассмотрены основные этапы разработки алгоритма распознавания. Показана нетривиальность задачи поиска оптимальных параметров классификаторов при многоклассовом распознавании. Предложен подход для формирования весовых коэффициентов классификаторов на основе вычисления «следа» ковариационной матрицы и дана его физическая трактовка.

The paper considers basic phases of recognition algorithm development. Non-triviality of task in searching optimum classifier parameters at many class recognition is shown. Approach for generating classifiers' weight coefficients is proposed based on calculating 'trace' of covariation matrix, and its physical handling is given.

В соответствии с принципами системности и иерархичности порядок разработки алгоритма распознавания целесообразно разбить на три взаимосвязанных этапа.

Первый этап заключается в определении полного множества признаков, характеризующих объекты классификации. В общем случае задание классов распознавания осуществляется посредством описания объекта распознавания многомерным вектором признаков

$$X = \{x_1, x_2, ..., x_N\},$$

где N – размерность вектора признаков.

В силу того что различные признаки имеют различную размерность, перед оценкой выборочных векторов средних и ковариационных матриц необходимо выполнить нормировку согласно выражению

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{\sigma_j}, i = 1,...,n_0 + n_1 + ... + n_M; j = 1,...,N,$$

где z_{ij} – нормированное значение j-го признака i-го объекта из обучающей выборки; x_{ij} – исходное значение j-го признака i-го объекта из обучающей выборки; $\overline{x}_j$ – выборочное среднее j-го признака; σ_j – среднеквадратическое отклонение j-го признака; n_k – объем обучающей выборки k-го класса, $k = 1,...,M$; M – число классов, участвующих в распознавании.

Повышения точности исходных данных можно добиться путем увеличения объемов обучающих наборов. Согласно центральной предельной теореме, при неограниченном увеличении n_k оценки и стремятся к своим действительным значениям [1]. Однако в реальных условиях объем выборки ограничен. Данное обстоя-

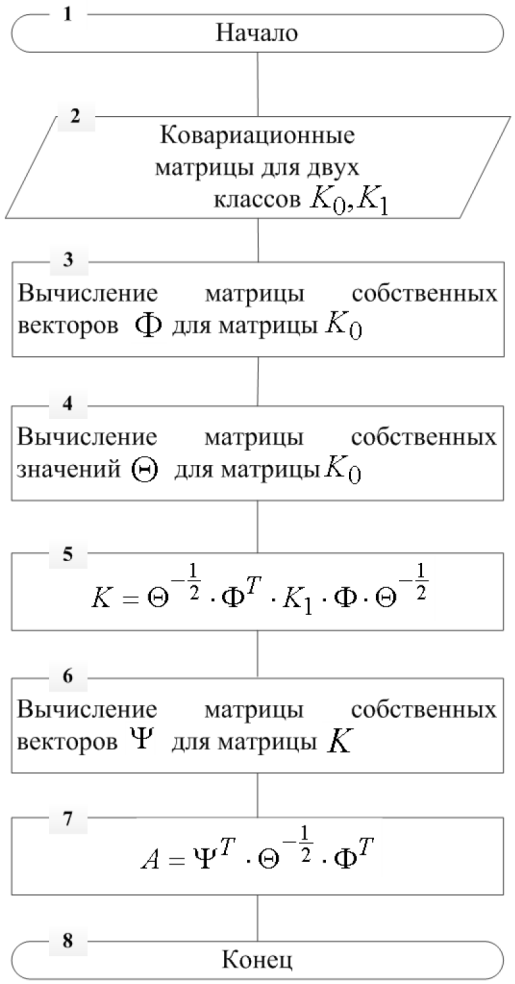

Рис. 1. Схема процедуры вычисления матрицы декоррелирующего преобразования

тельство приводит к необходимости проведения декорреляции исходных данных [2].

Схема процедуры вычисления матрицы декоррелирующего преобразования для двух классов представлена на рис. 1.

Таким образом, после выполнения преобразования

$$\widetilde{X} = A \cdot X$$

компоненты вектора $\widetilde{X}$ оказываются независимыми.

На рис. 2–3 представлены примеры поведения двумерного признакового пространства для двух классов до и после применения декоррелирующего преобразования.

Декоррелирующее преобразование приводит к статистической независимости параметров, формирующих признаковое пространство, что позволяет уменьшить количество ложных тревог и пропусков цели.

Второй этап разработки алгоритма распознавания заключается в определении решающего правила, в соответствии с которым признаковое пространство (пространство наблюдений) разбивается на гиперплоскости (гиперсферы), соответствующие количеству классов (гипотез). В случае многоклассовой классификации наиболее эффективным средством решения такого типа задач является метод опорных векторов (МОВ).

Основная идея МОВ состоит в получении границы сферической формы вокруг набора классифицируемых данных с минимальным радиусом и максимальным количеством объектов обучающей выборки внутри нее [3].

Для определения принадлежности образа X объекта к классу используется следующее выражение (решающее правило в общем виде):

$$f(X) = R^2 - \|X - c\|^2 \geq 0, \qquad (1)$$

где $\overline{c}$ – центр гиперсферы, R – радиус гиперсферы.

На третьем этапе в соответствии с выбранным критерием принятия решения осуществляется определение порога принятия решения.

Качество сформированного правила принятия решения (1) полностью определяется вероятностями ошибочных решений, допускаемых при распознавании объектов. Ошибка распознавания возникает из-за пере-

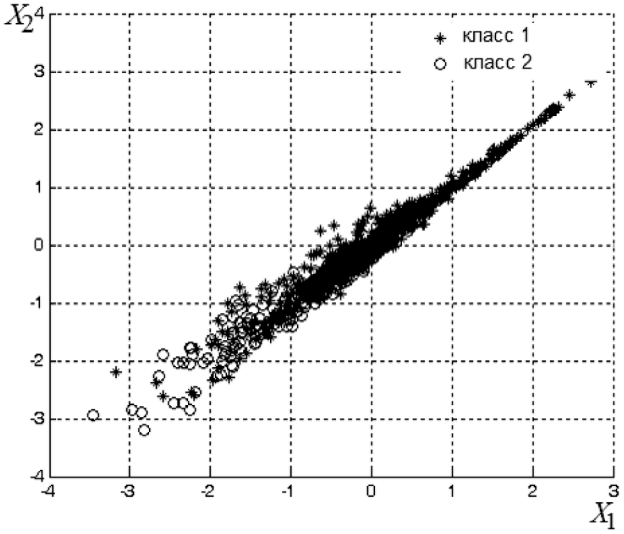

Рис. 2. Признаковое пространство двух классов до применения декоррелирующего преобразования

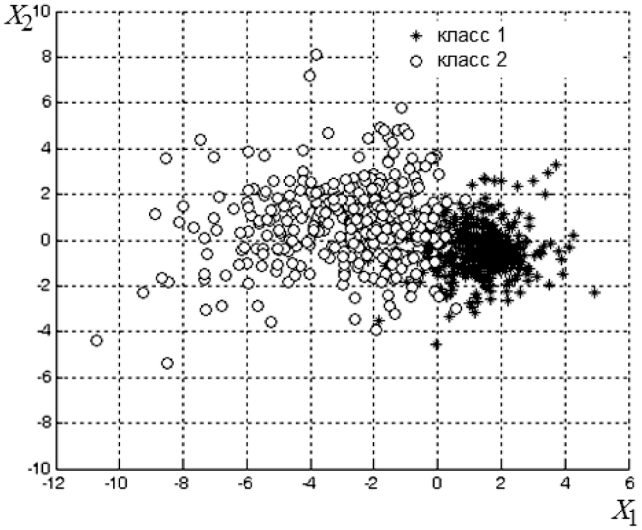

Рис. 3. Признаковое пространство двух классов после применения декоррелирующего преобразования

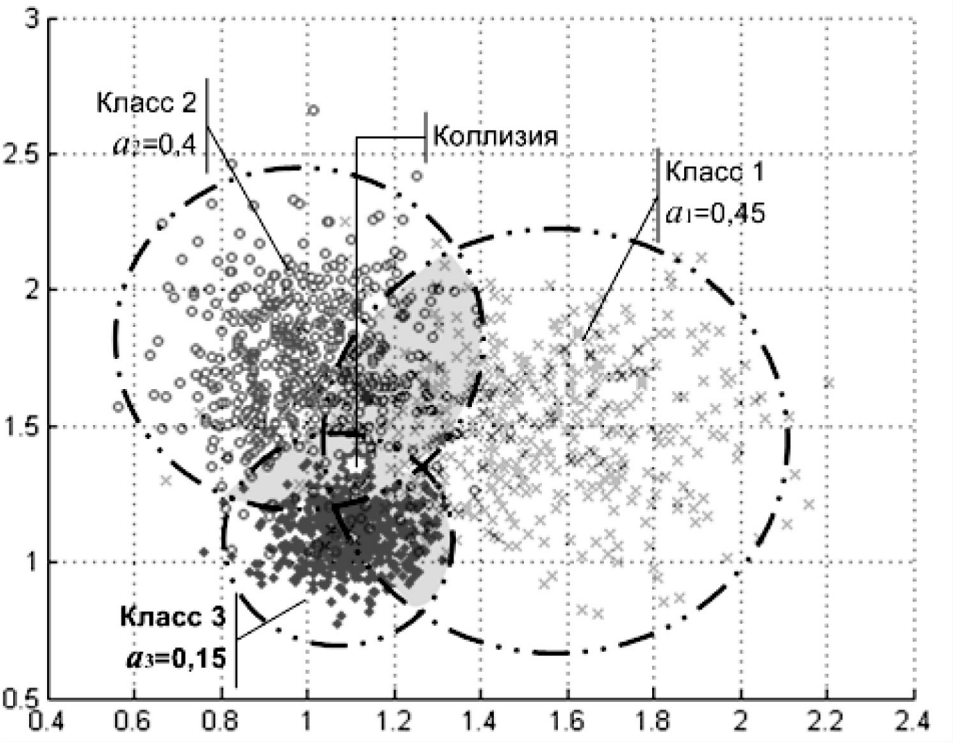

Рис. 4. Графическая интерпретация коллизии

сечения классов в многомерном признаковом пространстве, когда объекты из обучающей выборки одного класса в результате проекции на гиперплоскость ошибочно относят к объектам другого класса. В случае отнесения объекта к нескольким классам, принятие решения о принадлежности вектора признаков к конкретному классу выполняется с использованием весовых коэффициентов значимости, полученных на этапе обучения классификаторов, за счет настройки параметров классификатора в процессе обучения [4].

Однако задача поиска оптимальных параметров классификаторов с точки зрения минимизации ошибки распознавания является нетривиальной, особенно в случае многоклассового распознавания.

В таких условиях расчет весовых коэффициентов предлагается производить по выборочным ковариационным матрицам [2] классов, участвующих в распознавании. Выражение для расчета ковариационной матрицы для k-го класса имеет следующий вид:

$$\hat{K}_k = \frac{1}{n_k} \cdot \sum_{i=1}^{n_k} (X_i - \hat{m}_k) \cdot (X_i - \hat{m}_k)^T \qquad (2)$$

где $\hat{m}_k$ – выборочные вектора средних. Расчет осуществляется в соответствии с выражением:

$$\hat{m}_k = \frac{1}{n_k} \cdot \sum_{i=1}^{n_k} X_i, \quad X_i^T = (x_{1i}, x_{2i}, ..., x_{Ni}) \qquad (3)$$

Выборочный вектор средних и выборочная ковариационная матрица являются состоятельными оценками, при этом вторая оценка, в отличие от первой, является смещенной, так как ее среднее значение

$$m\{\hat{K}_k\} = \left[\frac{n_k}{n_k - 1}\right] \cdot K_k . \qquad (4)$$

Следовательно, для нахождения несмещенной оценки $\hat{K}_k$ необходимо умножить (2) на поправочный коэффициент $\frac{n_k}{n_k - 1}$:

$$\hat{K}_k = \frac{1}{n_k - 1} \cdot \sum_{i=1}^{n_k} (X_i - \hat{m}_k) \cdot (X_i - \hat{m}_k)^T . \qquad (5)$$

Таким образом, матрицу можно представить в следующем виде:

$$\hat{K}_k = \begin{pmatrix} \sigma_{11}^2 & \mathrm{cov}(x_1, x_2) & \cdots & \mathrm{cov}(x_1, x_N) \\ \mathrm{cov}(x_2, x_1) & \sigma_{22}^2 & \cdots & \mathrm{cov}(x_2, x_N) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{cov}(x_N, x_1) & \mathrm{cov}(x_N, x_2) & \cdots & \sigma_{NN}^2 \end{pmatrix}. \qquad (6)$$

Диагональными элементами ковариационной матрицы являются дисперсии отдельных признаков σ_{ii}^2, а недиагональными элементами – ковариации пары значений i-го и j-го признаков.

Вычислив сумму диагональных элементов матрицы $tr\,K_k$ («след» матрицы), возможно оценить качество признакового пространства для k-го классификатора. При этом полученную оценку целесообразно использовать при формировании значений весовых коэффициентов, что можно представить в следующей аналитической форме записи:

$$a_k = \frac{tr\,K_k}{tr\,K_1 + tr\,K_2 + ... + tr\,K_M},\ k = 1,\ldots, M,$$

где a_k – весовой коэффициент k-го классификатора; $tr\,K_k$ – след ковариационной матрицы k-го классификатора; M – количество классификаторов.

Физический смысл получаемых коэффициентов состоит в том, что при возникновении коллизий решение принимается в пользу классификатора с наименьшим значением весового коэффициента (рис. 4). Под коллизией понимается отнесение анализируемого объекта к двум и более классам.

Таким образом, предложенная процедура формирования значений весовых коэффициентов является инвариантной к параметрам классификатора, что позволяет снизить ошибку распознавания при невозможности оптимальной настройки классификаторов.

Литература

1. Гмурман, В. Е. Теория вероятностей и математическая статистика / В.Е. Гмурман. – М.: "Высшая школа", 1972. – 368 с.

2. Фукунага, К. Введение в статистическую теорию распознавания образов / К. Фукунага. – М.: Наука. Главная редакция физико-математической литературы, 1979. – 368 с.

3. Tax, D.M.J. Support Vector Data Description / D.M.J. Tax, R.P.W. Duin // Machine Learning. – 2004. – No. 54. – P. 45–66.

4. Субботенко, А. В. Алгоритм поэкземплярного распознавания на основе метода опорных векторов / А.В. Субботенко // Информация и Космос. – 2014. – № 2. – С. 47– 53.

5. Мерков, А. Б. Распознавание образов: Построение и обучение вероятностных моделей / А.Б. Мерков. – Ленанд, Москва, 2014. – 240 с.

6. Фомин, Я. А. Распознавание образов: теория и применения / Я.А. Фомин. – 2-е изд. – Москва: ФАЗИС, 2012. – 429 с.