МЕТОДИ МАТЕМАТИЧНОГО МОДЕЛЮВАННЯ

Итерационный метод построения векторного классификатора

ШУМЕЙКО А.А., СОТНИК С.Л.

Днепродзержинский государственный технический университет, Iveonik Systems

Статья посвящена построению метода классификации с использованием итерационного алгоритма вычисления критерия качества на основе расстояния Маханолобиса.

Стаття присвячена побудові метода класифікації з використанням ітераційного алгоритму вирахування критерію якості на основі відстані Маханолобиса.

Article is dedicated to building of the method to classification with use iteration algorithm of the calculation criterion quality on base of the Mahalanobis distance.

В различных областях человеческой деятельности (экономике, финансах, медицине и др.) постоянно возникает необходимость решения задач выявления скрытых зависимостей и поддержки принятия оптимальных решений. Как правило, точные методы отстают от потребностей реальной жизни. Здесь требуются универсальные, простые и надежные подходы, в основе которых могут быть использованы технологии и подходы математической теории распознавания и классификации. Данные подходы в качестве исходной информации используют выборки описаний-наблюдений объектов, предметов, ситуаций или процессов (выборки прецедентов), при этом каждое отдельное наблюдение-прецедент записывается в виде вектора числовых значений отдельных его свойств-признаков. Выборки признаковых описаний являются обычно первичными исходными данными, которые повседневно возникают в различных предметных областях. Для решения такого рода задач разработано достаточно много методов (см.[1],[2]), однако быстро растущие объемы обрабатываемой информации вынуждают использовать для построения методов классификации использовать новые методы, основанные на самых различных конструкциях ([3],[4]). В данной работе предлагается использовать итерационный алгоритм преобразования Карунена-Лоева для упрощения вычислительной схемы при использовании классификатора на основе расстояния Махаланобиса.

1. Постановка задачи. Будем использовать следующую модель задачи классификации.

Ω – множество объектов распознавания (пространство образов).

$\omega\in\Omega$ объект распознавания (образ).

$g(\omega):\Omega\to\Re$, $\Re=\{1,2,\ldots,n\}$- индикаторная функция, разбивающая пространство образов Ω на n непересекающихся классов $\Omega^1,\Omega^2,...,\Omega^n$. Индикаторная функция неизвестна наблюдателю.

X – пространство наблюдений, воспринимаемых наблюдателем (пространство признаков).

$x(\omega):\Omega\to X$ – функция, ставящая в соответствие каждому объекту ω точку $x(\omega)$ в пространстве признаков. Вектор $x(\omega)$ – это образ объекта, воспринимаемый наблюдателем.

В пространстве признаков определены непересекающиеся множества точек $\Xi[i]\subset X$ i=1,2...,n, соответствующих образам одного класса.

$\varphi(x)$: $X\to\Re$ решающее правило - оценка для $g(\omega)$ на основании $x(\omega)$, т.е. $\varphi(x)=\varphi(x(\omega))$.

Пусть $x_\nu = x(\omega_\nu), \nu = 1,2,...,N$ доступная наблюдателю информация о функциях $g(\omega)$ и $x(\omega)$, но сами эти функции наблюдателю неизвестны. Тогда $(g_\nu, x_\nu), \nu = 1,2,...,N$ — есть множество прецедентов.

Задача заключается в построении такого решающего правила $\varphi(x)$, чтобы распознавание проводилось с минимальным числом ошибок.

Обычный случай – считать пространство признаков евклидовым, а качество решающего правила измеряют частотой появления правильных решений. Обычно его оценивают, наделяя множество объектов Ω, некоторой вероятностной мерой.

На нынешний момент наиболее распространенным является байесовский подход, который исходит из статистической природы наблюдений. За основу берется предположение о существовании вероятностной меры на пространстве образов, которая либо известна, либо может быть оценена. Цель состоит в разработке такого классификатора, который будет правильно определять наиболее вероятный класс для пробного образа. Тогда задача, состоит в определении "наиболее вероятного" класса, Задано n классов $\Omega_1,\Omega_2,\ldots,\Omega_n$, а также $P(\Omega_i \mid x)$ – вероятность того, что неизвестный образ, представляемый вектором признаков x, принадлежит классу Ω_i. $P(\Omega_i \mid x)$ называется апостериорной вероятностью, поскольку задает распределение индекса класса после эксперимента (a posteriori – т.е. после того, как значение вектора признаков x было получено). Естественно выбрать решающее правило таким образом: объект относим к тому классу, для которого апостериорная вероятность выше. Такое правило классификации по максимуму апостериорной вероятности называется

Байесовским. Таким образом, для Байесовского решающего правила необходимо получить апостериорные вероятности $P(\Omega_i \mid x)$. Формула Байеса, полученная Т. Байесом в 1763 году, позволяет вычислить апостериорные вероятности событий через априорные вероятности и функции правдоподобия. Пусть $\Omega_1, \Omega_2, \ldots, \Omega_n$ – полная группа несовместных событий – $\bigcup_{i=1}^{n} \Omega_i = \Omega, \Omega_i \bigcap_{i \neq j} \Omega_j = \varnothing$. Тогда апостериорная вероятность имеет вид:

$$P(\Omega_i \mid x) = \frac{P(\Omega_i) P(x \mid \Omega_i)}{\sum_{i=1}^{n} P(\Omega_i) P(x \mid \Omega_i)},$$

где $P(\Omega_i)$ – априорная вероятность события Ω_i, $P(x \mid \Omega_i)$ – условная вероятность события x при условии, что произошло событие Ω_i.

Если известно или с достаточным основанием можно считать, что плотность распределения функций правдоподобия $P(x \mid \Omega_i)$ является гауссовской. то применение классификатора Байеса приводит к тому, что образы, характеризующиеся нормальным распределением проявляют тенденцию к группированию вокруг среднего значения, а их рассеивание пропорционально среднеквадратическому отклонению. Для случая многих переменных Гауссова плотность имеет вид

$$p(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2} \left((x-\mu) \Sigma^{-1} (x-\mu)^T \right)^2 \right),$$

где

$$\Sigma = \begin{vmatrix} E[(x_1-\mu_1)(x_1-\mu_1)] & \ldots & E[(x_n-\mu_n)(x_1-\mu_1)] \\ \vdots & \ddots & \vdots \\ E[(x_1-\mu_1)(x_n-\mu_n)] & \ldots & E[(x_n-\mu_n)(x_n-\mu_n)] \end{vmatrix}$$

ковариационная матрица. Здесь $x = (x_1, x_2, \ldots x_n)$ проверяемый объект, $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ – математическое ожидание соответствующего класса. Заметим, что на диагонали ковариационной матрицы стоят значения дисперсий.

Соответствующая мера

$$\|x - \Omega\|^2_{\Sigma^{-1}} = (x-\mu) \Sigma^{-1} (x-\mu)^T.$$

называется расстоянием Махаланобиса (Mahalanobis). В случае, если все события класса независимы, то все коэффициенты ковариационной матрицы, кроме стоящих на диагонали, будут равны нулю. Таким образом, евклидово расстояние является частным случаем расстояния Махаланобиса.

Использование расстояния Махаланобиса ограничивается существенными ограничениями – для того, чтобы корреляционная матрица была невырождена, необходимо, чтобы количество признаков было не меньше количества элементов класса, что для реальных задач далеко не всегда, выполнимо. Во-вторых, степень вырождения корреляционных матриц больших размерностей делает вычисление обратной матрицы неустойчивым.

2. Преобразование Карунена-Лоева. Пусть имеется n векторов $X^0, X^1, \ldots, X^{n-1}$. Требуется построить m векторов $Y^0, Y^1, \ldots, Y^{m-1}$ так, чтобы восстановление по этому множеству давало наименьшую среднеквадратичную ошибку восстановления по m векторам. Таким образом, нужно найти минимум величины

$$\varepsilon\left(\left\{ \alpha_k^{\mu}(m) \right\}_{\mu=0\ k=0}^{m-1\ n-1}, \left\{ Y^{\mu} \right\}_{\mu=0}^{m-1} \right) = \sum_{k=0}^{n-1} \left\| X^k - \sum_{\mu=0}^{m-1} \alpha_k^{\mu}(m) Y^{\mu} \right\|_2^2 \quad (1)$$

по всем множествам Y^{μ} и векторам $\alpha_k^{\mu}(m)$ таким, что $\sum_{k=0}^{n-1} (\alpha_k^{\mu}(m))^2 = 1$.

Введя матричные обозначения

$$A = \left\{ \alpha_k^{\mu}(m) \right\}_{\mu=0\ k=0}^{m-1\ n-1},\ \mathrm{Y} = \left\{ Y^{\mu} \right\}_{\mu=0}^{m-1},\ \mathrm{X} = \left\{ X^{\mu} \right\}_{\mu=0}^{n-1},$$

рассмотрим задачу

$$\varepsilon(A, Y) = \|X - AY\|_2^2 \to \min \quad (2)$$

по всем A, Y. Здесь $B^2 = tr\, B^T B = tr\, B B^T$ (tr – след матрицы, то есть сумма элементов главной диагонали, что равно сумме собственных значений матрицы).

Будем полагать, что матрицы A и Y невырождены и rang(A) = rang(Y) = m.

Заметим, что если rang(X) = m, то существует точное представление

$$X^k = \sum_{\mu=0}^{m-1} \alpha_k^{\mu}(m) Y^{\mu}$$

для $k = 0,1,\ldots,n-1$.

Имеет место следующее утверждение.

Теорема. Если rang(X) ≥m, то минимум ε (A, Y) достигается в том случае, когда строки матрицы Y являются собственными векторами ковариационной матрицы $X^T X$, которые соответствуют m максимальным собственным значениям, кроме того, A = XY^T и обе матрицы A и Y ортогональны.

Для полноты изложение приведем доказательство этого утверждения. Так как функция цели представляет собой квадратичный функционал, то необходимые и достаточные условия задачи (2) будут иметь вид

$$\begin{cases} \dfrac{\partial \varepsilon(A,Y)}{\partial A} = (X - AY) Y^T = 0, \\ \dfrac{\partial \varepsilon(A,Y)}{\partial Y} = A^T (X - AY) = 0. \end{cases}$$

Поскольку матрицы A и Y невырождены, то

$$\begin{cases} A = XY^T (YY^T)^{-1}, \\ Y = (A^T A)^{-1} A^T X. \end{cases} \quad (3)$$

Матрица YY^T симметричная, невырожденная, положительно определенная, следовательно, существует невырожденная матрица M такая, что

$$M^{-1} Y Y^T (M^{-1})^T = I,$$

где I единичная матрица.

Кроме того, матрица $M^T A^T AM$ симметричная и невырожденная, тогда существует ортогональная матрица U такая, что

$$U^T M^T A^T A M U = diag(\lambda_1, \lambda_2, \ldots, \lambda_m) \equiv \Lambda$$

– диагональная матрица.

Полагая $V = MU$, рассмотрим $\hat{A} = AV$ и $\hat{Y} = V^{-1} Y$, тогда

$$\hat{A}\hat{Y} = AVV^{-1}Y = AY,$$
$$\hat{A}^T\hat{A} = U^T(M^TA^TAM)U = \Lambda,$$
$$\hat{Y}\hat{Y}^T = U^{-1}(M^{-1}YY^T(M^{-1})^T = (U^{-1})^T = (U^TU)^{-1} = I.$$

Первое равенство означает, что $\hat{A}$ и $\hat{Y}$ удовлетворяют необходимому условию экстремума и соотношение (3) в силу диагональности матрицы, $\hat{A}^T\hat{A}$ и $\hat{Y}\hat{Y}^T$ примет вид

$$\begin{cases}\hat{A} = X\hat{Y}^T,\\ \Lambda\hat{Y} = \hat{A}^TX.\end{cases}$$

Из первых двух соотношений сразу получаем $\Lambda\hat{Y} = \hat{Y}\hat{X}^T\hat{X}$, то есть строки матрицы $\hat{Y}$ являются собственными векторами матрицы $\hat{X}^T\hat{X}$, а диагональные элементы $\lambda_1, \lambda_2, \ldots, \lambda_m$ – соответствующие им собственные значения. Тогда

$$\varepsilon(\hat{A},\hat{Y}) = \left\|X - \hat{A}\hat{Y}\right\|_2^2 =$$
$$tr(X^T - \hat{Y}^T\hat{A}^T)(X - \hat{A}\hat{Y}) =$$
$$= trX^T(X - \hat{A}\hat{Y}) = trX^TX - trX^T\hat{A}\hat{Y} =$$
$$X^2 - tr\hat{Y}^T\Lambda\hat{Y} =$$
$$= X^2 - tr\Lambda = \sum_{k=0}^{n-1}\lambda_k - \sum_{k=0}^{m-1}\lambda_k = \sum_{k=m}^{n-1}\lambda_k,$$

таким образом, минимум $\varepsilon(\hat{A},\hat{Y})$ достигается в случае, если $\lambda_1, \lambda_2, \ldots, \lambda_m$ наибольшие из n собственные значения матрицы X^TX.

Так как $A^TA = \Lambda$, то $Y^0, Y^1, \ldots, Y^{m-1}$ (главные значения) не коррелируются, поэтому это преобразование называют декоррелирующим или преобразованием Карунена-Лоэва.

Описанный метод определения главных компонент является достаточно ресурсоемким и неустойчивым, особенно в случае, если собственные значения матрицы близки к нулю.

Для наших целей более эффективным является использование итерационного метода определения главных компонент. Для этой цели рассмотрим задачу (1) с другой точки зрения.

Для случая m = 1 задача (1) сводится к определению одной компоненты Y⁰, которая наилучшим образом восстанавливает все исходные данные X

$$\varepsilon\left(\left\{\alpha_k^0\right\}_{k=0}^{n-1}, Y^0\right) = \sum_{k=0}^{n-1}\left\|X^k - \alpha_k^0Y^0\right\|_2^2 \to \min \qquad (4)$$

по всем $\left\{\alpha_k^0\right\}_{k=0}^{n-1}, Y^0$ при условии

$$\sum_{k=0}^{n-1}(\alpha_k^0)^2 = 1. \qquad (5)$$

Если $\left\{\tilde{\alpha}_k^0\right\}_{k=0}^{n-1}$ *и* $\tilde{Y}^0$ есть решение этой задачи и

$$\tilde{X}_0 = \left\{X^k - \tilde{\alpha}_k^0\tilde{Y}^0\right\}_{k=0}^{n-1}$$

– ошибка восстановления данных первой главной компонентой, то решая задачу

$$\varepsilon\left(\left\{\alpha_k^1\right\}_{k=0}^{n-1}, Y^1\right) = \sum_{k=0}^{n-1}\left\|\tilde{X}^k - \alpha_k^1Y^1\right\|_2^2 \to \min \qquad (6)$$

по всем $\left\{\alpha_k^1\right\}_{k=0}^{n-1}, Y^1$ при условии

$$\sum_{k=0}^{n-1}(\alpha_k^1)^2 = 1,$$

получаем вторую главную компоненту $\tilde{Y}^1$ и соответствующий вектор $\left\{\tilde{\alpha}_k^1\right\}_{k=0}^{n-1}$ и т.д.

При фиксированных $\left\{\alpha_k^0\right\}_{k=0}^{n-1}$ задача (4) решается методом наименьших квадратов, а в силу того, что функция цели представляет собой квадратичный функционал, необходимое и достаточное условия экстремума совпадают. Таким образом, решение задачи сводится к поиску решения уравнения

$$\frac{\partial\varepsilon\left(\left\{\alpha_k^0\right\}_{k=0}^{n-1}, Y^0\right)}{\partial Y^0} = -2\sum_{k=0}^{n-1}(X^k - \alpha_k^0Y^0)\alpha_k^0 =$$
$$= -2\left(\sum_{k=0}^{n-1}X^k\alpha_k^0 - \sum_{k=0}^{n-1}(\alpha_k^0)^2Y^0\right) = 0.$$

Отсюда получаем

$$Y^0 = \frac{\sum_{k=0}^{n-1}X^k\alpha_k^0}{\sum_{k=0}^{n-1}(\alpha_k^0)^2},$$

учитывая условие (5), имеем

$$Y^0 = \sum_{k=0}^{n-1}X^k\alpha_k^0.$$

Следующий шаг будем делать, исходя из предположения, что в задаче (4) нам известна компонента Y^0 и требуется найти экстремум по $\left\{\alpha_k^0\right\}_{k=0}^{n-1}$

$$\frac{\partial\varepsilon\left(\left\{\alpha_k^0\right\}_{k=0}^{n-1}, Y^0\right)}{\alpha_v^0} = -2(X^v - \alpha_v^0Y^0)Y^0 =$$
$$= -2\left(\left\langle X^v, Y^0\right\rangle - \alpha_v^0\left\langle Y^0, Y^0\right\rangle\right) = 0,$$

то есть

$$\alpha_v^0 = \frac{\left\langle X^v, Y^0\right\rangle}{\left\langle Y^0, Y^0\right\rangle},$$

где, как обычно, $\langle X, Y\rangle$ – скалярное произведение векторов X и Y.

Далее, считая, найденные $\left\{\tilde{\alpha}_k^0\right\}_{k=0}^{n-1}$ известными, повторяем весь процесс, пока не произойдет стабилизация ошибки. Полученные Y^0 будем считать первой главной компонентой.

Применяя этот алгоритм к ошибке восстановления, находим вторую главную компоненту и т.д.

Подробный алгоритм вычисления главных компонент выглядит следующим образом:

Итак, пусть имеется n компонент $X^0, X^1, \ldots, X^{n-1}$. Требуется построить m доменов $Y^0, Y^1, \ldots, Y^{m-1}$ так, чтобы восстановление по этим доменам давало наименьшую среднеквадратичную

ошибку восстановления по m доменам. Таким образом, нужно найти минимум величины

$$\sum_{k=0}^{n-1}\left\| X^k - \sum_{\mu=0}^{m-1}\alpha_k^{\mu}(m)Y^{\mu}\right\|_2 \qquad (7)$$

по всем множествам Y^{μ} и векторам $\alpha_k^{\mu}(m)$ таким, что

$$\sum_{k=0}^{n-1}(\alpha_k^{\mu}(m))^2 = 1.$$

Числа $\alpha_k^{\mu}(m)$, полученные в результате решения этой задачи от m не зависят, что, в конечном счете, позволяет свести к m задачам вида

$$\min\left\{\sum_{k=0}^{n-1}\left\| X^k - \alpha_k^{\mu}Y^{\mu}\right\|_2 \left| Y^{\mu}, \alpha_k^{\mu} : \sum_{k=0}^{n-1}(\alpha_k^{\mu})^2 = 1\right.\right\} \qquad (8)$$

Рассмотрим итерационный процесс, приводящий к решению задачи (8), а заодно и задачи (7).

Пусть, вначале, i = 0 и $\alpha_k(0) = \frac{1}{\sqrt{n}}$. Вычислим

$$Y(i) = \sum_{k=0}^{n-1}\alpha_k(i)X^k \qquad (9)$$

далее вычислим числа $\alpha_k^*(i) = \left\langle Y(i), X^k\right\rangle$ и проведем нормировку, то есть

$$\alpha_k(i+1) = \frac{\alpha_k^*(i)}{\sqrt{\sum_{k=0}^{n-1}(\alpha_k^*(i))^2}}.$$

Полагая $i := i+1$, продолжим итерационный процесс N раз, где N таково, что стабилизирует либо домен $Y(i)$, либо значения $\alpha_k(i)$. Как правило, для этого достаточно использовать 10-20 итераций. После этого полагаем $Y^0 = Y(N)$ и $\overline{\alpha}_k^0 = \alpha_k(N)$. Ясно, что ошибка восстановления каждой компоненты будет равна

$$\Delta X^k = X^k - \tilde{X}^k,$$

где $\tilde{X}^k = \overline{\alpha}_k^0 Y^0$ восстановление k-й компоненты по домену Y^0.

Полученную ошибку восстановления будем воспринимать как компоненту, к которой (в качестве исходных данных) повторим тот же итерационный процесс, то есть полагаем i = 0 и $\alpha_k(0) = \frac{1}{\sqrt{n}}$. Вычисляем $Y(i) = \sum_{k=0}^{n-1}\alpha_k(i)\Delta X^k$ и числа

$$\alpha_k^*(i) = \left\langle Y(i), \Delta X^k\right\rangle.$$

После нормировки получаем

$$\alpha_k(i+1) = \frac{\alpha_k^*(i)}{\sqrt{\sum_{k=1}^{n-1}(\alpha_k^*(i))^2}}.$$

Полагая $i := i+1$ продолжим итерационный процесс. После стабилизации итерационного процесса получаем $Y^1 = Y(i)$ и $\alpha_k^1 = \alpha_k(i)$. Далее находим $\Delta\tilde{X}^k = \alpha_k^1 Y^1$ и к полученной ошибке восстановления $\Delta^2 X^k = \Delta X^k - \Delta\tilde{X}^k$ итерационно применим тот же процесс n раз.

При достаточном числе итераций $\Delta^n X^k = 0$ для всех k, то есть алгоритм реализует полное разложение компонент X^k по доменам $Y^k (k = 0,1,\ldots,n-1)$.

Восстановление по m доменам будет равно

$$X_m^v = \sum_{k=0}^{m-1}\alpha_v^k Y^k .$$

Приведем основные свойства главных компонент

1. Имеет место равенство

$$\min\left\{\sum_{k=0}^{n-1}\left\| X^k - \sum_{\mu=0}^{m-1}\alpha_k^{\mu}(m)Y^{\mu}\right\|_2 \left| Y^{\mu}, \alpha_k^{\mu}(m) : \sum_{k=0}^{n-1}(\alpha_k^{\mu}(m))^2 = 1\right.\right\} =$$

$$= \sum_{k=0}^{n-1}\left\| X^k - \sum_{\mu=0}^{m-1}\alpha_k^{\mu}Y^{\mu}\right\|_2 .$$

2. Если

$$\hat{X}^v = \sum_{k=0}^{n-1}\alpha_v^k Y^k ,$$

то имеет место равенство Парсеваля

$$\sum_{v=0}^{n-1}\left\|\hat{X}^v\right\|_2^2 = \sum_{v=0}^{n-1}\left\|Y^v\right\|_2^2 ,$$

3. Более того, для $m = 1,\ldots,n$ и

$$\hat{X}_m^v = \sum_{k=0}^{m-1}\alpha_v^k Y^k$$

выполняется соотношение

$$\sum_{v=0}^{m}\left\|\hat{X}_m^v - \hat{X}^v\right\|_2^2 = \sum_{v=m+1}^{n-1}\left\|Y^v\right\|_2^2 .,$$

4. Векторы $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$ образуют ортонормированную систему.
5. При этом $\left\|Y^v\right\|_2^2 (v = 0,\ldots,n-1)$ собственные числа корреляционной матрицы $\left[\left\langle X^i X^j\right\rangle\right]$, векторы $\alpha_v (v = 0,\ldots,n-1)$ ее собственные векторы.
6. Последние частотные домены содержат “белый шум”.

3. Применение итерационного алгоритма преобразования Карунена-Лоева к построению векторного классификатора. Итак, пусть даны n векторов $X^0, X^1, \ldots, X^{n-1}$ и m главных компонент $Y^0, Y^1, \ldots, Y^{m-1}$ вместе с векторами $\alpha_k^{\mu}(m) k = 0,\ldots,m-1, \mu = 0,\ldots,n-1$.

Вычислим расстояние между вектором Z и множеством $\left\{X^i\right\}_{i=0}^{n-1}$. В соответствии с метрикой Маханалобиса

$$\left\| Z - \left\{X^i\right\}_{i=0}^{n-1}\right\|_{\Sigma^{-1}} =$$

$$= \sqrt{\left(Z - E\left(\left\{X^i\right\}_{i=0}^{n-1}\right)\right)\Sigma^{-1}\left(Z - E\left(\left\{X^i\right\}_{i=0}^{n-1}\right)\right)^T},$$

где Σ ковариационная матрица.

Замечая, что для главных компонент ковариационная матрица такова, что на ее диагонали стоят $\left\|Y^v\right\|_2^2 (v = 0,\ldots,m-1)$, остальные равны нулю, получаем

$$\Sigma^{-1}=\begin{pmatrix} \frac{1}{\left\|Y^0\right\|_2^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\left\|Y^1\right\|_2^2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\left\|Y^m\right\|_2^2} \end{pmatrix}$$

Кроме того, если

$$\hat{X}=\sum_{v=0}^{n-1}\hat{X}^v=\sum_{v=0}^{n-1}\sum_{k=0}^{m-1}\alpha_v^k Y^k=\sum_{k=0}^{m-1}\sum_{v=0}^{n-1}\alpha_v^k Y^k=$$
$$=\sum_{k=0}^{m-1}\Lambda^k Y^k,$$

где

$$\Lambda^k=\sum_{v=0}^{n-1}\alpha_v^k\ ,$$

Тогда орт центрального вектора множества будет равен

$$E\left(\left\{X^i\right\}_{i-0}^{n-1}\right)=\frac{\hat{X}}{\left\|\hat{X}\right\|_2}.$$

Для определения расстояния нужно получить разложение проверяемого вектора по базису главных направлений

$$\hat{Z}=\sum_{k=0}^{m-1}\beta^k Y^k,\ \text{где}\ \ \beta^k=\frac{\left\langle Z,Y^k\right\rangle}{\left\langle Y^k,Y^k\right\rangle}.$$

В результате может измениться нормировка, поэтому, нужно нормировать единицей.

Таким образом, расстояние Махаланобиса между проверяемым вектором и заданным множеством, будет иметь вид

$$\left\|Z-\left\{X^i\right\}_{i=0}^{n-1}\right\|_{\Sigma^{-1}}^2=$$
$$=\sum_{k=0}^{m-1}\frac{1}{\left\|Y^k\right\|_2^2}\left(\frac{\beta^k Y^k}{\left\|\sum_{k=0}^{m-1}\beta^k Y^k\right\|_2}-\frac{\Lambda^k Y^k}{\left\|\sum_{k=0}^{m-1}\Lambda^k Y^k\right\|_2}\right)^2=$$
$$=\sum_{k=0}^{m-1}\left(\frac{\beta^k}{\left\|\sum_{k=0}^{m-1}\beta^k Y^k\right\|_2}-\frac{\Lambda^k}{\left\|\sum_{k=0}^{m-1}\Lambda^k Y^k\right\|_2}\right)^2$$

что существенно упрощает вычисления и, кроме того, делает процедуру нахождения расстояния Махаланобиса более устойчивой.

Выводы

Результаты работы показали, что использование итерационного алгоритма для построения преобразования Карунена-Лоева позволяет построить эффективные алгоритмы классификации, основанные на расстоянии Махаланобиса.

ЛИТЕРАТУРА

1. Tou J., Gonzalez R. Recognition principles: Addison-Wesley Publishing Company, 1974.
2. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining.- СПб.: БХВ-Петербург, 2004.- 336 с.
3. Шумейко А.А., Сотник С.Л., Лысак М.В. Использование генетических алгоритмов в задачах классификации текстов// Тези доповідей на VI міжнародній науково-практичній конференції «Математичне та програмне забезпечення інтелектуальних систем» (MPZIS-2008), Дніпропетровськ, 2008.- С. 345-346.
4. Berry Michael W., Browne Murray Lecture notes in data mining.-World Scientific Publishing Co, Pte, Ltd.: Singapore, 2006, 222 p.
5. Gorban A. N., Kegl B., Wunsch D., Zinovyev A. Y. (Eds.), Principal Manifolds for Data Visualization and Dimension Reduction, Series: Lecture Notes in Computational Science and Engineering 58, Springer, Berlin - Heidelberg - New York, 2007, XXIV, 340 p.

пост. 23.04.09