УДК 004.021:004.85

СЕ ЦЗИНЬБАО (КНР)

МОДЕРНИЗИРОВАННЫЙ АЛГОРИТМ КЛАССИФИКАЦИИ НА ОСНОВЕ МЕТОДА ОПОРНЫХ ВЕКТОРОВ

An improved algorithm based on support vector machine for text classification is considered in this paper. Experiment results show that improved algorithm based on support vector machine is better for its speed and classification accuracy than known support vector machine method.

Использование метода опорных векторов SVM (Supporting Vector Machines) относится к числу наиболее распространенных и успешных способов решения задач распознавания образов [1–3]. В настоящее время одной из проблем, возникающих при обучении SVM, является выбор внутренних параметров алгоритма, т. е. таких параметров, которые задает сам пользователь и не изменяемых при обучении [4]. Цель данной работы заключается в разработке метода, позволяющего, используя принцип наименьшего евклидова расстояния, автоматически подбирать значения параметров границ решения в SVM.

Метод опорных векторов в задачах классификации

Рассматривается задача обучения по прецедентам [X, Y], где X – пространство объектов, Y – множество ответов. Требуется построить алгоритм $X \to Y$, аппроксимирующий целевую зависимость на всем пространстве X. Признаковый набор обучающих данных представляет собой набор пар $(x_1, y_1), (x_2, y_2), ..., (x_s, y_s)$, $x_i \in R^n$, $y_i \in \{1, -1\}$, $i = 1, 2, ..., s$. Необходимо найти правила классификации $C(x)$, позволяющие наилучшим образом распознавать один из двух классов.

Задача нахождения оптимального разделения множества векторов на два класса может выполняться с помощью линейной решающей функции. Тогда разделяющей два класса гиперплоскостью будет

$$(\omega \cdot x) + b = 0,$$

где ω – вектор весовых коэффициентов, b – некоторое число. Значения ω и b находим как решения оптимизационной задачи квадратичного программирования, минимизируя для обучающей выборки функцию

$$\varphi(\omega) = \frac{1}{2}\|\omega\|^2 = \frac{1}{2}(\omega \cdot \omega)$$

при ограничениях

$$y_i[(\omega \cdot x_i) + b] \geq 1, \;\; i = 1, 2, ..., s,$$

что согласно условиям Куна – Таккера равносильно поиску седловой точки лагранжиана

$$L(\omega, b, \alpha) = \frac{1}{2}(\omega \cdot \omega) - \sum_{i=1}^{n} \alpha_i \left\{ y_i \left[(\omega \cdot x_i) + b \right] - 1 \right\},$$

причем $\alpha_i > 0$ – коэффициенты лагранжиана. Для неопорных векторов соответствующие α_i равны 0, так что расчет выполняется только для опорных векторов.

В том случае, когда классы линейно не разделимы, применяются следующие подходы.

Первый (см., например, [4]) основан на возможности ошибок в обучающей коллекции. Вводится множество специальных переменных $c_i \geq 0,$ характеризующих величину ошибки. К структуре минимизируемой функции добавляется штраф

$$\varphi(\omega) = \frac{1}{2}(\omega \cdot \omega) + R \sum_{i=1}^{s} c_i,$$

а ограничения принимают вид

$$y_i[(\omega \cdot x_i) + b] \geq 1 - c_i, \;\; i = 1, 2, ..., s.$$

Настраиваемый параметр C позволяет регулировать отношения между двумя целями – максимизациией ширины разделяющей полосы и минимизацией суммарной ошибки.

Альтернативный подход основан на замене в приведенных формулах скалярного произведения нелинейной функцией ядра, т. е. скалярным произведением в пространстве с большей размерностью. Размерность получаемого пространства может быть больше размерности исходного, преобразование, сопоставленное скалярному произведению, будет нелинейным, а, значит, функция, соответствующая в исходном пространстве оптимальной разделяющей гиперплоскости, также будет нелинейной. В этом пространстве может существовать оптимальная разделяющая гиперплоскость.

Вне зависимости от используемого подхода одной из основных проблем остается трудоемкость процедуры обучения. Известен ряд модификаций классического SVM, например предложенный в [5] ESVM (Editing Supporting Vector Machines), основанный на следующем алгоритме.

Шаг 1. Выполнить предварительное обучение с помощью SVM, получить границы решения.

Шаг 2. Удалить выборки, которые находятся в некоторой области вблизи границы решения, и неправильные выборки.

Шаг 3. Выполнить обучение с помощью SVM для новых обучающих выборок, чтобы определить новые границы решения.

Шаг 4. В случае необходимости редактировать исходные выборки, используя новые границы решения, удалить неправильные выборки, получить еще раз новые обучающие выборки, а затем обновить границы решения.

Модернизированный метод опорных векторов

В настоящей статье предлагается модернизированный SVM, основанный на том, что в некоторой области вблизи границы решения по знаку класса каждой выборки и его ближайшего соседа определяются новые границы решения. В отличие от [5] этот подход очень прост, и результаты тестирования показывают, что по сравнению с SVM скорость и точность классификации улучшаются.

Предлагается следующая стратегия для выборок обучения: в некоторой области вблизи границы решения найти ближайшего соседа каждой точки. Затем каждую точку, если она похожа на своего ближайшего соседа, необходимо сохранить. Если точка с ближайшим соседом является гетерогенной, то удалить точку. По знаку класса каждой выборки и его ближайшего соседа определить новые границы решения.

Используя евклидово расстояние в качестве расстояния между двумя векторами x_i и x_j,

$$x_i = (x_i^1, x_i^2, ..., x_i^n), \;\; x_j = (x_j^1, x_j^2, ..., x_j^n),$$

определим его как

$$D(x_i, x_j) = \sqrt{\sum_{p=1}^{n} (x_i^p - x_j^p)^2},$$

и расстояние до ближайшего соседа будем находить по приведенной формуле.

Рассмотрим теперь алгоритм, реализующий этот метод. Признаковый набор обучающих данных $(x_1, y_1), (x_2, y_2), ..., (x_s, y_s)$, $x_i \in R^n$, $y_i \in \{1, -1\}$, $i = 1, 2, ..., s$. Обучающие выборки представлены в виде матрицы

$$TR_{s\times(n+1)} = [XY], \quad X = \begin{bmatrix} x_1 \\ \vdots \\ x_s \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_s \end{bmatrix}.$$

Алгоритм модернизированного SVM:

Шаг 1. Вычислить для каждой точки расстояния до остальных точек, расстояние до себя положить равным ∞.

for p=1 to s

$\{ Z_{1\times s} = (z_{ij}), z_{ij} = \infty, i = 1, j = 1, 2, ..., s;$

for q=1 to s

{if $q \neq p$, $z_{1q} = D(x_p, x_q);\}$

}

Шаг 2. Выполнить поиск ближайшего соседа.

$NN_{s\times 1} = (nn_{ij}), nn_{ij} = 1, i = 1, 2, ..., s, j = 1$

$t = 1; value = z_{11};$

for q=1 to s

{if $z_{1q} < value\{value = z_{1q}; t = q;\}$

$nn_{p1} = t;\}$

Шаг 3. Определение знака класса каждого вектора согласовать с его ближайшими соседями, отнесенными к категории 1 и –1

$L_{s\times 1} = (l_{ij}), l_{ij} = 1, i = 1, 2, ..., s, j = 1;$

for p=1 to s

{if $y_p \neq y_{nn_p}, l_{p1} = -1;\}$

Шаг 4. Удалить гетерогенный вектор, который несовместим с классом ближайшего соседа

$i = 0;$

for p=1 to s

{ if $l_{(p-i)1} = -1$

{ Удаление *p-i* строки в матрице *TR* и *L*;

i=*i*+1;}

}

После этих 4 шагов доступны модернизированные обучающие выборки *TR*.

Результаты экспериментальных исследований

В вычислительном эксперименте были использованы пакет программного обеспечения Matlab SVM [6] и база данных Ringnorm, содержащая 7400 20-мерных выборок из 2 классов.

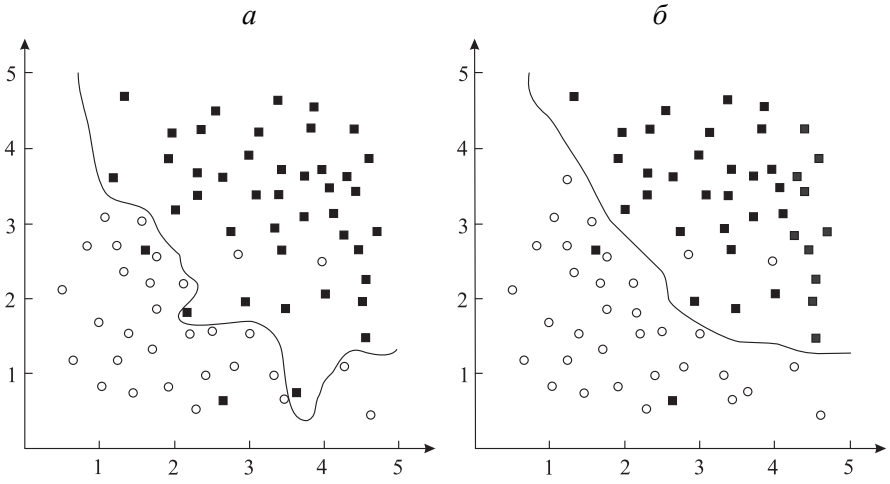

Рис. 1. Границы решения задачи классификации обычного (*а*) и модернизированного SVM (*б*)

В качестве ядра функции было взято гауссовское ядро с параметрами σ = 0,5 и C = 100. Для обучения использовались 4000 выборок, остальные 3400 – для тестирования модернизированного SVM.

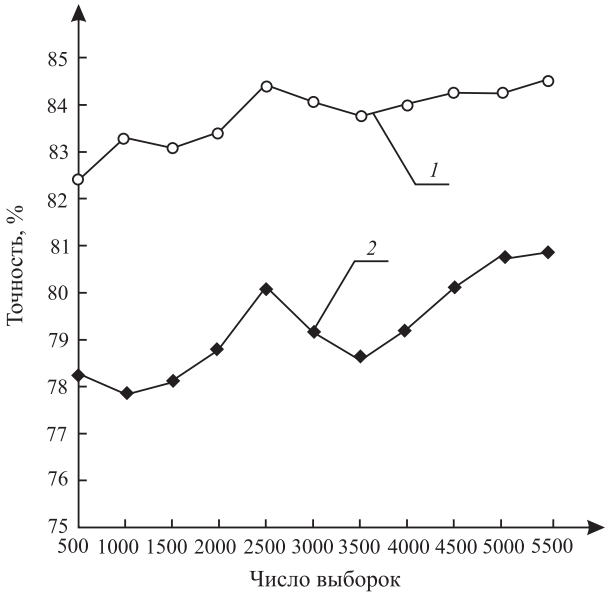

Рис. 2. Сравнение точности модернизированного SVM (*1*) с SVM для разных объемов чисел выборки (*2*)

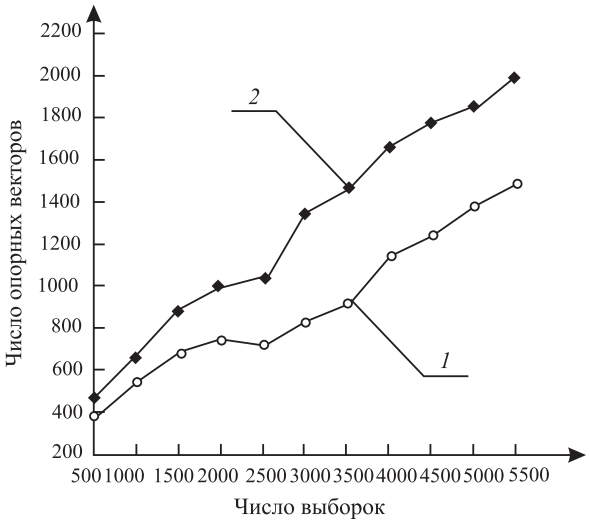

Рис. 3. Сравнение числа опорных векторов модернизированного SVM (*1*) с SVM для разных объемов чисел выборки (*2*)

На рис. 1 показано, что граница решения классификации модернизированного SVM проще, чем SVM. Точность классификатора модернизированного SVM выше по сравнению с SVM (рис. 2), в то время как число его опорных векторов меньше, чем у SVM (рис. 3).

Таким образом, предложенная модификация SVM позволяет улучшить точность классификации, сократив при этом время классификации за счет уменьшения числа опорных векторов.

1. Vapnik V.N. // IEEE Transactions on Neural Networks. 1999. 10 (5). P. 988.
2. Vapnik V.N. The Nature of Statistical Learning Theory. 2nd ed. New York, 1999.
3. Zhang Xue-Gong // Acta Automatica Sinica. 2000. 26 (1). P. 32.
4. Zhaoqi Bian, Zhang Xue-Gong. Pattern Recognition. Beijing, 2000.
5. Ke Hai-Xin, Zhang Xue-Gong // Proceedings of International Joint Conference on Neural Networks. Washington, 2001. 2. P. 1464.
6. http://svm.first.gmd.de/

Поступила в редакцию 24.12.10.

Се Цзиньбао – аспирант кафедры телекоммуникаций и информационных технологий. Научный руководитель – кандидат физико-математических наук, доцент, заведующий кафедрой кибернетики Ю.И. Воротницкий.