УДК 004.021

КЛАССИФИКАЦИЯ ДАННЫХ МЕТОДОМ ОПОРНЫХ ВЕКТОРОВ

И.И. ИВАНОВ
(Представлено: Е.С. ГАТИХО)

Анализируется классификация данных методом опорных векторов: описываются причины использования, описание алгоритма, его плюсы и минусы. Применение метода опорных векторов позволяет создать высокоточные системы классификации данных, имеющих место в статистическом анализе и методах машинного зрения.

Классификация данных – одна из центральных задач машинного обучения. В данном направлении интенсивно применяются методы оптимизации и аналитической геометрии. Задача классификации данных состоит в нахождении ответа на вопрос, к какому из определенных классов будут принадлежать новые наблюдаемые данные, основываясь на уже имеющемся наборе данных, классы которых известны. Каждый объект данных представляется как вектор (точка) в n-мерном пространстве (упорядоченный набор n чисел). Каждая из этих точек принадлежит только одному из определенных классов. Вопрос состоит в том, можно ли разделить точки гиперплоскостью размерности $n - 1$. Это типичный случай линейной разделимости. Искомых гиперплоскостей может быть множество, поэтому полагают, что максимизация зазора между классами способствует более уверенной классификации. То есть, можно ли найти такую гиперплоскость, чтобы расстояние от нее до ближайшей точки было максимальным. Это эквивалентно тому, что расстояние между двумя ближайшими точками, лежащими по разные стороны гиперплоскости, максимально. Если такая гиперплоскость существует, она называется оптимальной разделяющей гиперплоскостью, а соответствующий ей линейный классификатор называется оптимально разделяющим классификатором [1].

Пример иллюстрации оптимальной и неоптимальных разделяющих плоскостей в рамках абстрактного набора данных, заданного множеством векторов размерности 2, приведен на рисунке 1.

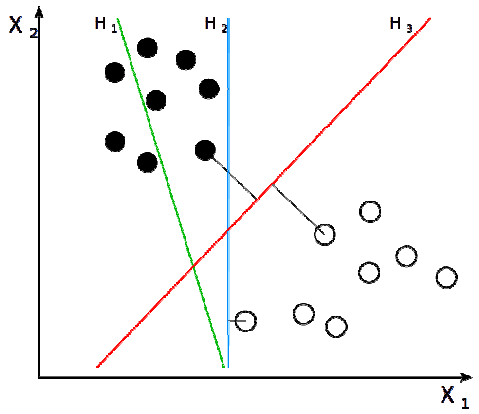

Рисунок 1. – Пример оптимальных и неоптимальных разделяющих плоскостей

При описании задачи полагается, что точки имеют вид:

$$\{(c_1 x_1),(c_2 x_2),,...,(c_p x_p)\},$$

где c_i принимает значение 1 или −1, в зависимости от того, какому классу принадлежит точка x_i.

Каждое x_i – это n-мерный вещественный вектор, обычно нормализованный значениями [0,1] или [–1,1]. Если точки не будут нормализованы, то точка с большими отклонениями от средних значений координат точек слишком сильно повлияет на классификатор. Можно рассматривать это как учебную коллекцию, в которой для каждого элемента уже задан класс, к которому он принадлежит.

Для того чтобы метод опорных векторов классифицировал их таким же образом, строится разделяющая гиперплоскость, имеющая вид:

$$\mathrm{w} \cdot x - b = 0.$$

Вектор w – перпендикуляр к разделяющей гиперплоскости.

Параметр $\frac{b}{\|w\|}$ равен по модулю расстоянию от гиперплоскости до начала координат.

Если параметр b равен нулю, гиперплоскость проходит через начало координат, что ограничивает решение. Так как производится поиск оптимального разделения, выделяются опорные векторы и гипер-

плоскости, параллельные оптимальной и ближайшие к опорным векторам двух классов. Можно показать, что эти параллельные гиперплоскости могут быть описаны следующими уравнениям (с точностью до нормировки):

$$\mathrm{w} \cdot x - b = 1,$$

$$\mathrm{w} \cdot x - b = -1.$$

Если обучающая выборка линейно разделима, то мы можем выбрать гиперплоскости таким образом, чтобы между ними не лежала ни одна точка обучающей выборки, и затем максимизировать расстояние между гиперплоскостями [2].

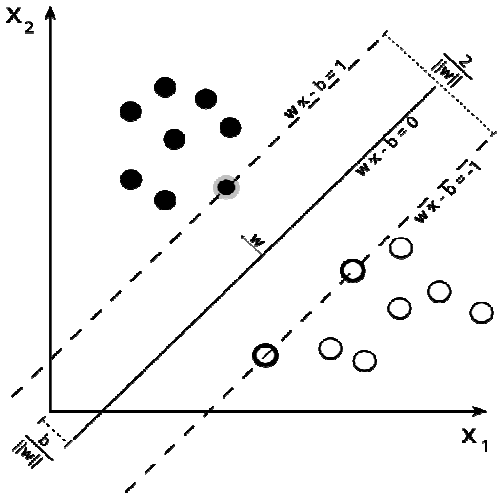

Рисунок 2. – Пример оптимальной гиперплоскости с выделенными опорными векторами

Ширину полосы между ними легко найти из соображений геометрии, она равна $\frac{2}{\|w\|}$. Таким образом, наша задача минимизировать $\|w\|$. Чтобы исключить все точки из полосы, мы должны убедиться для всех *i*, что и

$$\mathrm{w} \cdot x_i - b \geq -1, \quad c_i = 1,$$

$$\mathrm{w} \cdot x_i - b \leq 1, \quad c_i = 1.$$

Это может быть также записано в виде:

$$c_i(\mathrm{w} \cdot x_i - b) \geq 1, 1 \leq i \leq n.$$

Метод опорных векторов обладает следующими преимуществами [2]:

- наиболее быстрый метод нахождения решающих функций;
- метод сводится к решению задачи квадратичного программирования в выпуклой области, которая всегда имеет единственное решение;
- метод находит разделяющую полосу максимальной ширины, что позволяет в дальнейшем осуществлять более уверенную классификацию;
- возможность классификации среди множества классов;
- малые временные затраты непосредственно на классификацию данных.

Однако, как и любой алгоритм, метод опорных векторов не лишен и своих недостатков, главным из которых является чувствительность к шумам и стандартизации данных.

Исходя из вышесказанного, можно сделать *вывод*, что применение метода опорных векторов позволяет создать высокоточные системы классификации данных, имеющих место в статистическом анализе и методах машинного зрения.

ЛИТЕРАТУРА

1. Support vector machine [Электронный ресурс]. – Режим доступа: https://en.wikipedia.org/wiki/Support_vector_machine. – Дата доступа: 02.09.2017.
2. Машина опорных векторов [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=SVM. Дата обращения: 02.09.2017.