



Международный журнал информационных технологий и энергоэффективности

Сайт журнала: <http://www.openaccessscience.ru/index.php/ijcse/>



УДК 004.421(043.3)

ИСПОЛЬЗОВАНИЕ МЕТОДА ОПОРНЫХ ВЕКТОРОВ В ЗАДАЧАХ КЛАССИФИКАЦИИ

Баев Н.О.

Филиал ФГБОУ ВО "НИУ "МЭИ" в г. Смоленске, Россия (214013, г. Смоленск, Энергетический проезд, дом 1); e-mail: arhol.dss@yandex.ru

Статья посвящена изложению использования метода опорных векторов в задачах классификации. В ходе анализа метода опорных векторов и сравнении его с другими методами было выяснено, что он является оптимальным для робототехнических систем. Также была выделена основная проблема метода – высокие требования к вычислительным ресурсам. Решение этой проблемы заключается в решении оптимизационной задачи Лагранжа генетическим алгоритмом, что сокращает время обучения системы.

Ключевые слова: метод опорных векторов, гиперплоскости, алгоритм обучения, классификация, генетический алгоритм, обучающая выборка.

USING THE METHOD OF SUPPORT VECTORS IN CLASSIFICATION TASKS

Baev N.O.

Smolensk Branch of the National Research University "Moscow Power Engineering Institute", Russia (214013, Smolensk, street Ehnergeticheskij, 1); e-mail: arhol.dss@yandex.ru

The article is devoted to the use of the method of support vectors in classification problems. During the analysis of the method of reference vectors and comparison with other methods, it was found that it is optimal for robotic systems. Also, the main problem of the method was identified - high demands on computational resources. The solution to this problem is to solve the optimization problem of Lagrange by a genetic algorithm, which reduces the learning time of the system.

Key words: method of support vectors, hyperplanes, learning algorithm, classification, genetic algorithm, training sample.

Задачи классификации широко используются в процессе выработки решений в ряде технических систем. Многие из них изначально должны строиться на принципах обучения (самообучения) [8,10]. К таким системам можно отнести робототехнические системы, системы поддержки принятия решений и другие. В этой связи задача построения эффективных алгоритмов обучения классификации является весьма актуальной.

В настоящее время известен ряд методов (алгоритмов) обучения задаче классификации [2,4,5]. Рассмотрим их достоинства и недостатки с точки зрения возможности применения в робототехнических системах (таблица 1).

Таблица 1 - Достоинства и недостатки методов обучения классификации

Метод	Достоинства	Недостатки
С 4.5	Простая реализация, интерпретация и отсутствие подготовки данных для их дальнейшего использования. Работа с категориальными и интервальными переменными. Использование модели «белого ящика», её оценка. Работа с большим объёмом информации.	Отсутствие оптимальности дерева решений в целом, необходимость регулировки его длины. Переизбыток данных и плохая читабельность.
К-means	Простота реализации, высокое быстродействие, работа с большим количеством данных.	Наличие неточностей в кластеризации, необходимость начального определения точного числа классов, чувствительность к выбору центров кластеров.
Метод опорных векторов (SVM)	Высокое быстродействие, единственно верное решение, нахождение максимальной ширины полосы разделения, вследствие чего производится уверенная классификация.	Большая чувствительность к шумам, стандартизации исходных данных, отсутствие общего подхода к автоматическому выбору ядра в случае линейно неразделимости классов.
CART	Непараметрический метод. Не нужно рассчитывать различные параметры вероятностного распределения. Отсутствие необходимости выбора переменных при анализе данных. Нечувствительность к шумам, высокое быстродействие.	Нестабильность дерева решений. Некорректное отображение деревьев со сложной структурой.

Как видно из данных, приведенных в таблице 1, весьма перспективным с точки зрения задачи обучения роботов является метод опорных векторов. Он позволяет проводить обучение в условиях поступления данных в реальном масштабе времени, обеспечивает уверенную классификацию и работает по малой выборке обучающих данных [3,7].

Проблемным вопросом применения этого метода в робототехнических системах являются достаточно высокие требования к вычислительным ресурсам. Рассмотрим возможность устранения этого недостатка и обеспечения высокого быстродействия метода.

Метод опорных векторов - это набор алгоритмов обучения с учителем, использующихся, в основном, для задач классификации. Принадлежит к семейству линейных классификаторов и может также рассматриваться как специальный случай регуляризации по Тихонову [1,6,9].

Каждый объект данных представляется как вектор (точка) в p -мерном пространстве. Каждая из этих точек принадлежит к одному из двух классов. Для разделения классов строят гиперплоскость размерности $(p-1)$. Искомых гиперплоскостей может быть много, алгоритм находит оптимальную, которая максимизирует зазор между классами, что способствует более уверенной классификации (рисунок 1).

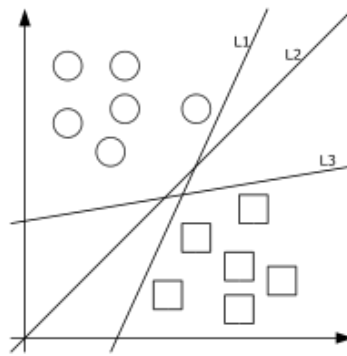


Рисунок 1 – Оптимальное разделение классов

На рисунке 1 представлено несколько классифицирующих разделяющих прямых (гиперплоскостей), из которых только одна (L_2) соответствует оптимальному разделению.

Классификация объектов в соответствии с методом опорных векторов реализуется в соответствии с зависимостью.

$$a(x) = \text{sign}(\sum_{j=1}^n w_j x^j - w_0) = \text{sign}(\langle w, x \rangle - w_0), \quad (1)$$

где $x = (x^1, \dots, x^n)$ - признаковое описание объекта x ;

w - вектор;

$w_0 \in R$ - параметр алгоритма;

$\langle w, x \rangle = w_0$ - разделяющая гиперплоскость в пространстве R^n .

Процесс построения разделяющей гиперплоскости представляет собой задачу квадратичного программирования, в конечном итоге сводится к поиску седловой точки функции Лагранжа, что приводит к задаче нелинейной оптимизации с ограничениями:

$$\left\{ \begin{array}{l} -L(\lambda) = -\sum_{i=1}^l \lambda_i + \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle \rightarrow \min_{\lambda}; \\ 0 \leq \lambda_i \leq C, i = 1, \dots, l; \\ \sum_{i=1}^l \lambda_i y_i = 0. \end{array} \right. \quad (2)$$

Важной особенностью метода опорных векторов является то, что он в обучающей выборке использует только часть – опорные векторы. Это свойство позволяет постепенно накапливать обучающие примеры и проводить обучение в режиме реального времени, что очень важно для роботов.

Учитывая небольшой размер обучающей выборки, и, следовательно, небольшое количество параметров в оптимизационной задаче, целесообразно рассмотреть возможность ее решения с помощью генетического алгоритма. Небольшое количество генов в особи обеспечит быструю сходимость алгоритма.

В начале алгоритма инициализируется входная выборка, содержащая набор переменных λ_i . Определяются классы принадлежности и координаты точек обучающей выборки. Затем происходит сравнение условий, необходимых для построения разделяющей гиперплоскости. После нахождения условий классификации происходит поиск лучшей хромосомы для построения разделяющей гиперплоскости. Процесс повторяется до тех пор, пока не будет достигнут минимум целевой функции.

Алгоритм обучения запускается при поступлении новых обучающих примеров. При этом, как правило, происходят небольшие уточнения положения разделяющей гиперплоскости, параметры которой определяются величинами λ_i . Это обстоятельство дает возможность формировать исходную популяцию генетического алгоритма на основе уже имеющихся параметров λ_i , что позволяет существенно сократить время поиска минимума оптимизируемой функции.

С учетом приведенных выше соображений можно предложить следующие этапы способа обучения:

- получение первичных обучающих примеров (для начала работы способа их должно быть по одному из каждого класса);
- решение задачи построения разделяющей гиперплоскости методом опорных векторов с помощью генетического алгоритма, сохранение параметров λ_i ;
- получение новых обучающих примеров, проверка примеров на предмет возможной принадлежности к множеству опорных векторов; если проверка дает отрицательный результат, то ожидание новых примеров; если – положительный, то реализация дообучения с помощью генетического алгоритма;
- в процессе дообучения исходная популяция генетического алгоритма формируется на основе уже имеющихся параметров λ_i ; в результате дообучения формируется новая разделяющая гиперплоскость;
- процесс обучения продолжается до принятия учителем решения о прекращении обучения.

Сравнение традиционного и предлагаемого способов обучения системы классификации показывает, что предлагаемый способ позволяет сократить время обучения на 5-8 %.

Главным преимуществом предлагаемого способа обучения на основе метода опорных векторов является простота его реализации по сравнению с другими подходами, приведёнными в таблице 1. Кроме того, снижаются требования к вычислительным ресурсам по сравнению с другими методами оптимизации – ещё один плюс применения генетического алгоритма. При этом обеспечивается высокое быстродействие способа за счет непрерывного уточнения оптимизируемых параметров и небольших размеров хромосом.

Список литературы

1. Арсенин В. Я. Методы математической физики и специальные функции. – М.: Наука, 1974.
2. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. – М.: Наука, 1979.
3. Вьюгин В. Математические основы машинного обучения и прогнозирования. – МЦМНО, 2014. – 304 с.
4. Журавлев Ю. И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. – М.: Фазис, 2006.
5. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. – Новосибирск: ИМ СО РАН, 1999.
6. Ильин В. А., Позняк Э. Г. Линейная алгебра. – М.: ФИЗМАТЛИТ, 2004.
7. Сафонов И., Сергей Завалишин - Частичное Обучение с Учителем на Небольших Исходных Выборках. – М.: Наука, 1988.
8. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. – М.: Финансы и статистика, 1989.
9. Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. – М.: Наука, 1979.
10. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. – Киев: Наукова думка, 2004.

References

1. Arsenin V. Ya. Methods of mathematical physics and special functions. - Moscow: Nauka, 1974.
 2. Vapnik V.N. Restoration of dependencies according to empirical data. - Moscow: Nauka, 1979.
 3. Vyugin V. Mathematical foundations of machine learning and forecasting. - MCM-NO, 2014. - 304 с.
 4. Zhuravlev Yu.I., Ryazanov V. V., Senko O. V. "Recognition". Mathematical methods. Software system. Practical applications. - M.: Phasis, 2006.
 5. Zagoruiko NG Applied methods of data and knowledge analysis. - Novosibirsk: IM SB RAS, 1999.
 6. Ilyin VA, Poznyak EG Linear algebra. - Moscow: FIZMATLIT, 2004.
 7. Safonov I., Sergey Zavalishin - Partial Training with the Teacher on Small Initial Samples. - Moscow: Nauka, 1978.
 8. Aivazyan SA, Buchstaber VM, Enyukov IS, Meshalkin LD Applied Statistics: Classification and Dimension Reduction. - Moscow: Finance and Statistics, 1989.
 9. Tikhonov AN, Arsenin V. Ya. Methods for solving ill-posed problems. - Moscow: Nauka, 1979.
 10. Schlesinger M., Glavach V. Ten lectures on statistical and structural recognition. - Kiev: Naukova Dumka, 2004.
-