УДК 004.021

ПРИМЕНЕНИЕ МЕТОДА ОПОРНЫХ ВЕКТОРОВ В ЗАДАЧАХ КЛАССИФИКАЦИИ ТЕКСТА ДЛЯ СИСТЕМ РАСПРЕДЕЛЕННОЙ ОБРАБОТКИ ИНФОРМАЦИИ

А.Г. ВОЛИК, А.Г. МУРЛИН

Кубанский государственный технологический университет,
350072, Российская Федерация, г. Краснодар, ул. Московская, 2

В статье описываются основы метода опорных векторов в задачах классификации текста. Описывается задача машинного обучения классификации текстов применительно к распределённым системам. Приводятся алгоритмы для классификации линейно разделимых и линейно неразделимых выборок
Ключевые слова: линейный, метод опорных векторов, гиперплоскость, линейно разделимый, линейно неразделимый

Классификация тестовых сообщений является одной из важных задач для обеспечения качественной работы социальных сервисов. Нежелательные сообщения рекламного характера, так называемый спам, приводят к ухудшению удобства использования и снижают удовлетворенность пользователей. В связи с этим выявление эффективных методов определения вида сообщений становится актуальной проблемой.

Классификация текстов представляет собой процесс присвоения текстам на естественном языке некоторых категорий с помощью различных алгоритмических методов. Обычно выделяют машинное обучение (machine learning) и информационный поиск (information retrieval) [1]. Машинное обучение представляет индуктивный процесс построения классификатора и требует заранее обработанной и классифицированной человеком выборки документов. Информационный поиск же заключается в определении человеком некоторого набора правил для оценки принадлежности документа к определенному классу. Одним способов решения данной задачи является метод опорных векторов.

Метод опорных векторов (support vector machine) представляет набор алгоритмов обучения с учителем, в основе которых лежит перевод исходных векторов в пространство более высокой размерности и поиск разделяющей

гиперплоскости с максимальным зазором в этом пространстве [2]. Данный подход подразумевает получение некоторого правила, позволяющего различать объекты двух классов.

В общем случае задача может быть описана следующим образом. Пусть имеется пространство объектов классификации $X = \{x_1 \dots x_n\} \in \mathbf{R}^n$, множество ответов $Y = \{y_1 \dots y_n\} \in \{-1; 1\}$, а также некоторая целевая зависимость $y^*: X \to Y$, значения которой известны только на ограниченной области объектов обучающей выборки $X^l = \{x_i, y_i\}_{i=1}^l$. Требуется построить алгоритм $a: X \to Y$, аппроксимирующий целевую зависимость на всем пространстве X [3].

Каждый объект классификации x_i представляет собой вектор (точку) в *n*-мерном пространстве. Каждая координата вектора определяет степень выраженности (наличия) некоторого признака у данного объекта. Число y_i, характеризует принадлежность соответствующего вектора к заданной категории (1 – принадлежит, -1 – в противном случае).

Простейшим способом решения данной задачи будет служить построение линейного порогового классификатора путем нахождения гиперплоскости, разделяющей два класса в пространстве $\mathbf{R}^n$, Уравнение (1) будет описывать гиперплоскость, разделяющую классы в данном пространстве.

$$w \cdot x_i = w_0 \qquad (1)$$

Таким образом, задача классификации сводится к нахождению вектора $w = \{w^1 \dots w^n\} \in \mathbf{R}^n$, и некоторого граничного значения w_0, такого, что:

$$a(x) = sign(w \cdot x_i - w_0) \qquad (2)$$

При этом вектор w будет перпендикулярен искомой гиперплоскости (вектор нормали), а занесение w_0 будет зависеть от кратчайшего расстояния между гиперплоскостью и началом координат [2].

В большинстве случае выделить единственную (оптимальную) разделяющую гиперплоскость не представляется возможным, и существуют другие положения, реализующие тоже разбиение выборки. Тогда можно выбрать такую разделяющую гиперплоскость так, чтобы она максимально далеко отстояла от ближайших к ней точек обоих классов. Это должно

способствовать более уверенной классификации, увеличивая зазор ε (margin) между объектами классов, тем самым приводя, линейный классификатор к классификатору с разделяющей полосой. Для этого случая, с учетом уравнения (1), получаем неравенство (3), задающее полосу разделения классов.

$$\varepsilon \leq w \cdot x_i - w_0 \leq \varepsilon \qquad (3)$$

Параметры линейного порогового классификатора (2) и неравенства (3) определены с точностью до нормировки и алгоритм $a(x)$ не изменится, если w и w_0 одновременно умножить на одну и ту же положительную константу. Удобно выбрать эту константу таким образом, чтобы для всех пограничных (т. е. ближайших к разделяющей гиперплоскости) объектов выполнялись условия (4).

$$w \cdot x_i - w_0 = y_i \qquad (4)$$

В качестве такого решения может служить выбор выбора $\varepsilon = 1$ и с

предварительным умножением неравенства на .

Таким образом, для всех векторов в обучающей выборке X^l следует (5), а границами полосы служат две параллельные гиперплоскости с направляющим вектором w [4].

$$w \cdot x_i - w_0 = \begin{cases} \leq -1, & \text{если } y_i = -1 \\ \geq 1, & \text{если } y_i = +1 \end{cases} \qquad (5)$$

Ширина получаемой разделяющей полосы равна на и чем она шире, тем более точным будет классификатор [3].

Возможны два случая обучающей выборки: линейно разделимая и линейно неразделимая.

В первом случае построение оптимальной разделяющей гиперплоскости сводится к задаче квадратичной оптимизации путем минимизации квадратичной формы при l ограничениях относительно n + 1 переменных w и w_0 (6).

$$\begin{cases} w \cdot w \to min; \\ y_i(w \cdot x_i - w_0) \geq 1, i = 1, \ldots, l \end{cases} \qquad (6)$$

Однако данный подход работает только для линейно разделимых выборок, а корме этого при наличии ошибки в обучающей выборке приводит к существенным разбросам положения получаемой гиперплоскости. Для улучшения алгоритма и смягчения реакции на ошибки в обучающей выборке ввести набор дополнительных переменных $\lambda_i \geq 0$ (двойственных переменных), характеризующих величину ошибки на объектах, что позволяет смягчить неравенства из (6).

$$y_i(w \cdot x_i - w_0) \geq 1 - \lambda_i, i = 1, \ldots, l \qquad (7)$$

При $\lambda_i = 0$, предполагается, что ошибка классификации отсутствует, а если $\lambda_i > 0$, то допускается ошибка. При объект попадает в разделительную полосу и считается относящимся к своему классу.

В итоге, с учетом (7) задача может быть сведена к решению системы (8).

$$\begin{cases} w \cdot w - C\sum_{i=1}^{l} \lambda_i \to min; \\ y_i(w \cdot x_i - w_0) \geq 1 - \lambda_i, i = 1, \ldots, l \\ \lambda_i \geq 0, i = 1, \ldots, l \end{cases} \qquad (8)$$

В системе (10) параметр C является управляющим параметром и позволяет регулировать отношение между шириной полосы и минимизацией суммарной ошибки [3].

По теореме Куна-Таккера эта задача эквивалентна двойственной задаче поиска седловой точки функции Лагранжа, необходимым которой является равенство нулю производных Лагранжиана, из которых вытекают соотношения (9) и (10) [3].

$$w = \sum_{i=1}^{l} \lambda_i y_i x_i \qquad (9)$$

$$\sum_{i=1}^{l} \lambda_i y_i \qquad (10)$$

Если $\lambda_i > 0$ и $w \cdot x_i - w_0 = y_i$, то объект обучающей выборки x_i называется опорным вектором (support vector).

С учетом последнего определения и подставляя (9) и (10) обратно в (8) можно прейти к эквивалентной задаче содержащей только двойственные переменные λ_i. Для (8) вектор w вычисляется по формуле (9), а для определения порога w_0 достаточно взять произвольный опорный вектор x_i и выразить w_0 из равенства $w_0 = w \cdot x_i - y_i$.

В итоге алгоритм классификации может быть записан в виде (11).

$$a(x) = sign\left(\sum_{i=0}^{t} \lambda_i y_i (x_i \cdot x) - w_0\right) \qquad (11)$$

Метод опорных векторов относится к семейству линейных классификаторов, использующихся для задач классификации и регрессионного анализа, одним из главных свойств которого является непрерывное уменьшение эмпирической ошибки классификации. Хотя само правило классификации в точности совпадает с моделью нейрона по МакКаллоку-Питтсу, критерий и методы настройки параметров в SVM радикально отличаются от персептронных (градиентных) методов обучения. При работе метода производится суммирование не по всей выборке, а только по опорным векторам, для которых $\lambda_i = 0$. Именно это свойство разреженности (sparsity) отличает SVM от других линейных разделителей – дискриминанта Фишера, логистической регрессии и однослойного персептрона.

ЛИТЕРАТУРА

1. Лифшиц Ю. Курс Алгоритмы для Интернета: Автоматическая классификация текстов [Электронный ресурс]. – Электрон. дан. – Режим доступа: http://logic.pdmi.ras.ru/~yura/internet/06ianote.pdf, свободный (дата обращения: 01.06.2014) – Загл. с экрана.

2. Лифшиц Ю. Курс Алгоритмы для Интернета: Метод опорных векторов (Support vector machines) [Электронный ресурс]. – Электрон. дан. – Режим доступа: http://logic.pdmi.ras.ru/%7Eyura/internet/07ianote.pdf, свободный (дата обращения: 01.06.2014) – Загл. с экрана.

3. Воронцов К. Лекция по методу опорных векторов [Электронный ресурс]. – Электрон. дан. – Режим доступа: http://www.ccas.ru/voron/download/SVM.pdf, свободный (дата обращения: 01.06.2014) – Загл. с экрана.

4. Drucker H., Wu D., Vapnik V. N. “Support Vector Machines for Spam Categorization” in “IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 10, NO. 5, SEPTEMBER 1999”. - NY: IEEE, 1999.

REFERENCES

1. Lifshits Y. Kurs Algoritmy dlya Interneta: Avtomaticheskaya klassifikatsiya tekstov. Retrieved from: http://logic.pdmi.ras.ru/~yura/internet/06ianote.pdf (Algorithms for Internet: Automated text classification)

2. Lifshits Y. Kurs Algoritmy dlya Interneta: Metod opornykh vektorov. Retrieved from: http://logic.pdmi.ras.ru/%7Eyura/internet/07ianote.pdf (Algorithms for Internet: Support vector machines).

3. Vorontsov K. Lektsiya po metodu opornykh vektorov. Retrieved from: http://www.ccas.ru/voron/download/SVM.pdf (Lection on Support vector machines).

4. Drucker H., Wu D., Vapnik V. N. “Support Vector Machines for Spam Categorization”. 1999. “IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 10, NO. 5, SEPTEMBER 1999”.

USING SUPPORT VECTOR MACHINES FOR TEXT CATEGORIZATION IN DISTRIBUTED INFORMATION PROCESSING SYSTEMS

A.G. VOLIK, A.G. MURLIN

Kuban State Technological University
2, Moskovskaya st., Krasnodar, Russian Federation, 350072

The article shows basic approaches of usage support vector machines for text categorization. Describes the problem of text categorization and machine learning in distributed systems. The algorithms for linearly separable and linearly nonseparable samples are shown.
Keywords: linear qualifier, support vector machines, hyperplane, nonseparable, separable