УДК 004.04.043; 004.912; 004.62

К.К. Духновская

Матрица взаимосвязи терминов во множестве информационных ресурсов для метода опорных векторов

Представлена стратегия повышения эффективности работы метода опорных векторов с применением ядра на основе матрицы взаимосвязи терминов во множестве информационных ресурсов.

The strategy of the efficiency improvement of the SVM method, using the kernel based on the matrix relationship of terms in the set of the information resources is presented.

Представлено стратегію підвищення ефективності роботи методу опорних векторів з застосуванням ядра на основі матриці взаємозв'язку термінів у множині інформаційних ресурсів.

Введение. Информационные ресурсы (ИР), которые подаются в электронных хранилищах глобальной или корпоративных сетей, служат базисом для принятия решений в государственных, научных, образовательных учреждениях, и определяют успешность их работы в целом. Интенсивный рост количества ИР, их доступность и высокая динамичность приводит к избытку информации. Для преодоления этих проблем создаются интегрированные банки ИР [1]. Во время формирования таких хранилищ проводится предварительная обработка ИР с целью их интеллектуального анализа. Одним из существенных этапов предварительной обработки ИР есть классификация.

В ряду современных методов классификации ведущее место занимает метод опорных векторов в силу своей строгой теоретической обоснованности. Данный метод применяют в теории распознавания образов, в интеллектуальном анализе данных (*Data Mining*), также он широко используется для построения поисковых систем на этапе классификации текстовых документов. Но алгоритмам из этого семейства свойственна проблема масштабируемости – большая ресурсоёмкость использования памяти и времени вычислений на этапе обучения. В данной статье представлена стратегия повышения эффективности работы метода опорных векторов с применением ядра на основе матрицы взаимосвязи терминов во множестве ИР. Вследствие такой стратегии увеличивается вес более информативных признаков и объединений этих признаков, что делает классификатор более быстрым и менее ресурсоемким.

Постановка задачи

Для данного исследования под информационным ресурсом будем понимать некоторое информационное сообщение, представленное в электронном виде. Любой информационный ресурс можно описать в виде кортежа:

$$D_j = < t_{j1}, t_{j2}, \ldots, t_{jM}, p_{j1}, p_{j2}, \ldots, p_{jn} >, \qquad (1)$$

где t_{ji} ($i = 1, ..., M$) – статистическая мера важности i-го термина j-го сообщения. Термины (понятия) – это имена мысленных образов, которые передаются в процессе обмена информацией. Термины содержатся в словаре, M – мощность словаря. Статистическая мера важности термина – это отношение частоты встречаемости термина в ИР к количеству всех терминов ИР. Остальные признаки в описании этого ИР обозначены как p_{ji} ($i = 1, \ldots, r$| r – количество дополнительных признаков), к ним могут относиться дата создания ИР, его автор, адрес, ссылки на другие ИР и пр. Вектор, представленный кортежем (1) называют *профилем* этого ресурса. Классификация ИР заключается в разбиении множества этих ресурсов на непересекающиеся группы с целью обеспечения минимального различия между ресурсами одной группы, соответствующей определенной содержательной тематике, и максимального различия между ресурсами других групп.

Пусть дано множество ИР: $D = \{D_j \mid D_j = <t_1, t_2, \ldots, t_M >\}$ и множество классов $Q = \{Q_k \mid k =$

$= 1 \dots N_c\}$. Каждый класс Q_i описывается некоторой структурой $F_k = \{D_{k1}, D_{k2}, ..., D_{kl}\}$. Процедура классификации f заключается в выполнении некоторых преобразований над ИР, после которых делается заключение о соответствии ресурса D_j одной из структур $F_k, f: D \to Q$.

Связь терминов в документе

Если пренебречь дополнительными признаками, то множество ИР может быть представлено следующим образом:

$$\begin{pmatrix} t_{1\,1} & \dots & t_{1\,m} \\ \dots & \dots & \dots \\ t_{n\,1} & \dots & t_{n\,m} \end{pmatrix}, \qquad (2)$$

где n – количество ИР в их множестве.

Из теории нечетких множеств связь двух терминов во множестве ИР может быть определена как

$$c_{ij} = \frac{\widetilde{n}_{ij}}{\widetilde{n}_i + \widetilde{n}_j - \widetilde{n}_{ij}}, \qquad (3)$$

где $\widetilde{n}_i$ – количество ИР, в которых встречается i-й термин; $\widetilde{n}_j$ – количество ИР, в которых встречается j-й термин; $\widetilde{n}_{ij}$ – количество ИР, в которых есть оба термина. Но в (3) не учитывается, с какой частотой содержатся термины в ИР. Коэффициент взаимосвязи c_{ij} может иметь одинаковое значение для терминов, которые несут основное содержание ИР и для несущественных терминов. Во избежание этого недостатка в (3) введем нормированную частоту термина в ИР t_{ij}:

$$c_{ij} = \frac{\frac{1}{2}\sum_{i_1=0}^{i_1=\tilde{n}_{ij}} (t_{ii_1} + t_{ji_1})}{\sum_{i_1=0}^{i_1=\tilde{n}_i} t_{ii_1} + \sum_{i_1=0}^{i_1=\tilde{n}_j} t_{ji_1} - \frac{1}{2}\sum_{i_1=0}^{i_1=\tilde{n}_{ij}} (t_{ii_1} + t_{ji_1})}. \qquad (4)$$

В результате матрица взаимосвязи терминов примет вид:

$$C = \begin{pmatrix} 1 & c_{12} & \dots & c_{1m} \\ c_{21} & 1 & \dots & c_{2m} \\ \dots & \dots & \dots & \dots \\ c_{m1} & c_{m2} & \dots & 1 \end{pmatrix}. \qquad (5)$$

По сути, это матрица корреляции двух терминов во множестве ИР. Ее коэффициенты указывают на статистическую зависимость частот двух терминов, при этом изменения значений одной или нескольких из этих величин приводят к систематическому изменению значений другой (других) величины.

Свойства матрицы взаимосвязи

Определение 1. Симметричная матрица A называется положительно определенной, если для любого x выполняется неравенство $x^T Ax > 0$ [2].

Свойство 1. Матрица взаимосвязи терминов C (5) во множестве ИР – положительно определенная.

Доказательство: Согласно критерию Сильвестра [2]: симметричная матрица положительно определена тогда и только тогда, когда главные ее миноры положительны. Доказательство строится на применении метода Гаусса и приведении матрицы (5) к треугольному виду (при этом учитывается, что матрица симметричная и ее элементы – положительные числа). При таких преобразованиях значения главных миноров не изменятся и будут равны произведению их диагональных элементов.

Свойство 2. Матрица взаимосвязи терминов во множестве ИР квадратична.

Утверждение 1. Для матрицы взаимосвязи терминов во множестве ИР существует такая матрица B, которая представляется как $B = C^{\frac{1}{2}}$.

Доказательство: из леммы Шура [2] следует утверждение, что если матрица C – симметричная, то существует ортогональная матрица S, столбцы которой есть собственные векторы матрицы C, и диагональная матрица V, элементы которой – собственные значения матрицы C, такие, что:

$$\mathrm{V} = \mathrm{S}^{-1} C S. \qquad (6)$$

Все собственные значения положительно определенной матрицы – положительные. Отсюда, все ненулевые элементы диагональной матрицы V больше нуля, а значит, существует $V^{\frac{1}{2}}$:

$$V = \begin{pmatrix} v_{11} & 0 \dots 0 & 0 \\ 0 & v_{ii} & 0 \\ 0 & \dots & v_{nn} \end{pmatrix},$$

$$B = V^{\frac{1}{2}} = \begin{pmatrix} \sqrt{v_{11}} & 0\ldots0 & 0 \\ 0 & \sqrt{v_{ii}} & 0 \\ 0 & 0\ldots0 & \sqrt{v_{nn}} \end{pmatrix}.$$

Если обе части выражения (6) слева умножить на S, а справа – на S^{-1} (это возможно, так как $SS^{-1} = S^{-1}S = E$ – единичная матрица), то:

$$S(S^{-1}CS)S^{-1} = SVS^{-1},$$

$$C = SVS^{-1} = SV^{\frac{1}{2}}V^{\frac{1}{2}}S^{-1} = SV^{\frac{1}{2}}(S^{-1}S)V^{\frac{1}{2}}S^{-1} =$$

$$= (SV^{\frac{1}{2}}S^{-1})(SV^{\frac{1}{2}}S^{-1}) = BB = B^2.$$

Определение 2. Функция $K : XxX -> R$ называется ядром, если она представляется в виде $K(x, x`)=[\varphi(x), \varphi(x`)]$ при некотором отображении $\varphi: X -> F$, где F-пространство со скалярным произведением [3].

Пусть $\varphi(d) = C^{\frac{1}{2}}d$, тогда ее ядро

$$K(d, d_1) = d^T C d_1. \qquad (7)$$

Утверждение 2. Функция, определяемая выражением (7), есть ядро.

Доказательство: $K(d, d_1) = d^T C d_1 = d^T \sqrt{C}^T \sqrt{C} d_1 =$

$$= \left[\varphi(d), \varphi(d_1)\right].$$

Свойство 3. Матрица взаимосвязи терминов во множестве ИР есть ядро.

Метод опорных векторов

Для решения задач классификации существует много подходов – вероятностный подход (например, наивный байесовский метод и его модификации), алгебраический подход (через различные меры близости ИР: евклидово расстояние и его модификации, манхэттенское расстояние**,** расстояние Махаланобиса и пр.). Метод опорных векторов сегодня – это метод классификации, результаты которого оцениваются как наиболее эффективные. Следует отметить, что по методу опорных векторов рассматривается задача бинарной классификации. Если во множестве ИР имеется большое число классов, то задача классификации может быть решена способом, при котором каждый класс отделяется от всех остальных. В этом случае каждая бинарная задача не зависит от других, и решать их можно параллельно на разных машинах.

Метод опорных векторов – это набор алгоритмов классификации типа *обучение с учителем*. Для реализации данного метода каждый ИР представляется как точка в M-мерном пространстве. Скоплениями этих точек будут определяться классы (рис. 1). Между этими классами проводят разделяющую гиперплоскость, т.е. строится такая гиперплоскость, чтобы расстояние между двумя ближайшими точками из разных классов было максимальным. Если такая гиперплоскость существует, то она называется *оптимальной разделяющей* гиперплоскостью [3].

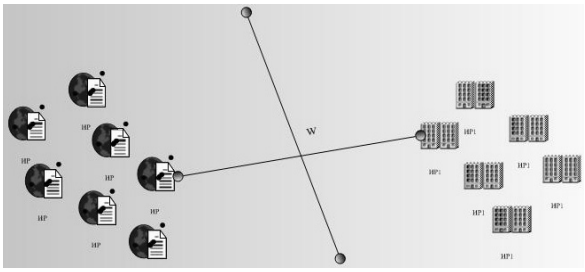

Рис. 1. Классификация методом опорных векторов

Пусть вектор w – опорный вектор: перпендикуляр к разделяющей гиперплоскости с точки класса; b – расстояние от гиперплоскости до начала координат (рис. 2). Тогда уравнение гиперплоскости имеет вид: $wx - b = 0$.

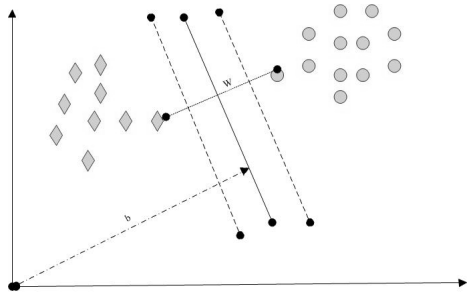

Рис. 2. Графическое представление метода опорных векторов

Относительно этой гиперплоскости все точки одного класса лежат по одну сторону. Если построить гиперплоскость, параллельную данной и проходящую через точку класса, ближайшую к оптимальной разделяющей гиперплоскости, то ее уравнение таково: $wx - b = 1$. Уравнением такой же гиперплоскости для другого класса будет $wx - b = -1$.

Между этими гиперплоскостями образуется полоса, которая должна быть свободна от точек одного и другого класса. Чтобы исключить все точки из полосы (см. рис. 2), необходимо проверить условие: $\begin{cases} wx_i - b \ge 1, k_i = 1 \\ wx_i - b \le -1, k_i = -1 \end{cases}$.

Проблема построения оптимальной разделяющей гиперплоскости сводится к задаче минимизации длины опорного вектора *w*. Это задача квадратичной оптимизации, которая представляется так [4]:

$$\begin{cases} \| w \|^2 \to \min \\ k_i (wx_i - b) \ge 1 \end{cases}. \quad (8)$$

Задача (8) – задача математического программирования. Если переписать ее в общем виде, то получим: $\begin{cases} f(x) \to \min \\ \varphi(x) \ge 0 \end{cases}$.

Для поиска решения такой задачи составляют функцию Лагранжа:

$$L(x, \lambda) = f(x) + \sum_{i=0}^{M} \lambda_i \varphi(x),$$

где λ_i – множители Лагранжа.

Согласно теореме Куна–Таккера [4] задача (8) примет вид:

$$\begin{cases} L(w, b, \lambda) = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{M} \lambda_i (k_i (wx_i - b) - 1) \to \min_{w,b} \max_{\lambda}, \\ \lambda_i > 0, \ 1 < i < M. \end{cases}$$

При этом она сводится к тождественной задаче, содержащей только двойственные переменные:

$$\begin{cases} -L(\lambda) = -\sum_{i=1}^{M} \lambda_i + \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \lambda_i \lambda_j k_i k_j [x_i, x_j] \to \min_{\lambda}, \\ \lambda_i > 0, \ 1 < i < M, \\ \sum_{j=1}^{M} k_i \lambda_j = 0. \end{cases}$$

Если задача решена, то *w* и *b* можно найти по формулам: $w = \sum_{i=1}^{M} \lambda_i k_i x_i$, $b = wx_i - k_i$.

В результате алгоритм классификации может быть записан в виде:

$$a(x) = \operatorname{sign}\left(\sum_{i=1}^{M} \lambda_i k_i [x_i, x] - b \right). \quad (9)$$

В модифицированных методах опорных векторов вместо скалярных произведений содержатся произвольные ядра [5], что позволяет избавиться от линейности. Заменяя скалярное произведение в (9) произвольным ядром, получают:

$$a(x) = \operatorname{sign}\left(\sum_{i=1}^{M} \lambda_i k_i K [x_i, x] - b \right). \quad (10)$$

Для более уверенной и эффективной классификации в формуле (10) в качестве ядра была взята матрица (5).

Решение двойственной задачи поиска седловой точки функции Лагранжа

Задача (10) с учетом и использованием ядра (5) принимает вид:

$$\begin{cases} -L(\lambda) = -\sum_{i=1}^{M} \lambda_i + \frac{1}{2} \sum_{i,j=1}^{M} \lambda_i \lambda_j k_i k_j d_j^T C d_i \to \min_{\lambda}, \\ \lambda_i > 0, \ 1 < i < M, \\ \sum_{i=1}^{M} k_i \lambda_j = 0. \end{cases} \quad (11)$$

Целевая функция данной задачи – квадратичная, а ограничение – линейными функциями, поэтому эту задачу относят к категории задач квадратичного программирования. К наиболее популярным методам решения задач данного класса относятся градиентные методы. Их применение в общем случае позволяет найти точку локального экстремума. Алгоритм решения задачи с помощью градиентных методов состоит в том, что, начиная с некоторой точки, осуществляется последовательный переход к другим точкам в сторону антиградиента до тех пор, пока не будет найдено допустимое решение исходной задачи. При нахождении решения задачи градиентными методами итерационный процесс продолжается, пока градиент функции в очередной точке не станет равным нулю или пока не будет превосходить какую-то бесконечно малую величину (точность полученного решения) [6].

Практическая реализация обучения данного метода классификации ИР может быть описана следующими этапами:

1. Оцифровка ИР: удаление различных управляющих знаков, тэгов, стоп-слов и представление тестового информационного сообщения в векторном виде.

2. Составление коэффициентов матрицы (5) для входящего множества ИР.

3. Начальное приближение: выбирается произвольный вектор, представляющий ИР одного класса, для которого ищется ближайший вектор другого класса. Для этого вектора ищем ближайший вектор из первого класса и т.д.

4. Решение задачи (11) методом градиентного спуска.

Применение метода опорных векторов отличается в лучшую сторону от других методов еще и тем, что данную задачу можно распараллелить. Учитывая гигантские мощности современных банков данных, размеры обучающей выборки должны оцениваться сотнями тысяч. При такой размерности применение стандартных численных методов квадратичного программирования становится невозможным. В настоящее время предложено несколько алгоритмов, оптимизирующих такие задачи. Один из них – метод последовательных оптимизаций (*SMO*) [7]. Согласно *SMO* на каждой итерации решается минимально возможная подзадача. Результатом такого разбиения будет много простых и независимых друг от друга подзадач, а значит, возможно их параллельное вычисление на разных машинах.

Качество классификации для метода опорных векторов на ядре матрицы взаимосвязи терминов во множестве документов

Работа классификатора может иметь два уровня ошибок. Ошибка *первого* уровня – если ИР ошибочно не расположен в нужном классе. Ошибки *второго* уровня – когда ИР ложно оказывается в определяемом классе. Пусть количество ИР в тестовом наборе равно N, из них N_p – количество правильно определенных классу ИР, а N_n – количество ИР, не имеющих отношения к классу. Тогда $N = N_p + N_n$. Пусть количество ложных пропусков – F_n, а ложных обнаружений – F_P, тогда количество корректных пропусков и обнаружений: $T_P=N_p–F_n$; $T_n = N_n – F_P$. Степень точности и полноты, часто используемые в задачах поиска информации, рассчитываются на основе характеристик T_P и F_P:

$$\text{precision} = P = \frac{T_p}{T_p + F_p} \times 100\%,$$

$$\text{recall} = R = \frac{T_p}{T_p + F_N} \times 100\% \text{ [8].}$$

Полнота измеряет долю корректной классификации по всем ИР данного класса. Точность измеряет долю правильных обнаружений всех выявленных ресурсов. Полнота и точность – величины, зависимые одна от другой. Во время разработки архитектуры классификатора ИР обычно приходится в качестве доминантной выбирать одну из двух характеристик. Если выбор пал на точность, это приводит к убыванию полноты из-за увеличения числа ложноположительных ответов. Рост полноты вызывает одновременное падение точности. Поэтому удобно для характеристики классификатора использовать одну величину, так называемую меру F_1, или меру Ван Ризбергена [8]: $F_1 = 2\dfrac{P \times R}{P + R}$.

Мера F_1 – одна из самых распространенных характеристик для подобного рода систем. В вычислении F_1 для задач классификации есть два основных подхода: суммарный F_1 **(**результаты по всем классам сводятся в одну таблицу, по которой затем вычисляется мера F_1) и средний F_1 (для каждого класса формируется собственное значение F_1, затем вычисляется среднее арифметическое для всех классов).

Процент ошибок можно определить метрикой правильности:

$$A(\text{accurancy}) = \frac{T_P + F_P}{N}.$$

В качестве рабочего материала для проведения экспериментов использована тестовая выборка ИР по двум научным дисциплинам: информационный поиск и механика сплошной среды, т.е. тестовую коллекцию нужно было разбить на два класса. В каждом из них было примерно одинаковое количество ИР – 200 и 220, что обеспечивало равномерность результатов – ни один из классов не выделялся только по количеству ИР в нем. Следует отметить, что точность определения ИР для отдельного класса может сильно зависеть от качества ресурсов именно этого класса. Общее количество ИР в выборке составило 420 документов. Они были разделены случайным образом на две

равные части по 210 ИР в каждой с сохранением примерно равного количества ресурсов по классам.

Обучение проводилось на одном из этих двух наборов, а тестирование – на другом. Далее все ИР из обучающей выборки разделили на пять частей. Изымая первую часть ИР из обучающей выборки, обучение классификатора проводилось на оставшихся 80 процентах ИР. Используя тестовую выборку, определялись показатели качества работы классификатора. Затем, изымая вторую часть из обучающей выборки, вычислялись другие значения показателей качества работы. В итоге было получено пять значений показателей качества работы классификатора. После этого наборы менялись местами и прогоны повторялись. В качестве окончательных результатов бралось их среднее арифметическое значение. Данное усреднение позволило сгладить результаты – тем самым сделать их более корректными.

Программная реализация данных методов классификации проводилась в среде *Embarcadero RAD Studio XE*2 на языке *C++*. Вследствие проделанной работы получены следующие результаты:

Метод классификации	**Точность**	**Полнота**	***F*1-мера**	**Процент ошибок**
Метод опорных векторов	72,9%	74,83%	73,85	0,49
Метод опорных векторов (ядро на основе матрицы связи терминов в коллекции ИР)	85,65%	75,06%	79,51	0,48

Заключение. В процессе анализа объективной оценки методов классификации имеется ряд проблем. Во-первых, для сравнения классификаторов все работы по тестированию методов классификации необходимо проводить в одинаковых условиях, т.е. на одних и тех же множествах ИР и классах, для которых известен корректный результат. Во-вторых, представление самих ИР в разных реализациях одного и того же метода может быть разным. На данном этапе проводятся работы по созданию стандартных коллекций текстов. К одной из наиболее популярных стандартных коллекций текстов относится *Reuters*-21578. Эта коллекция состоит из новостных сообщений по экономике, спорту и культуре. К сожалению, коллекция *Reuters* представляет тексты на английском языке, а в рамках данного проекта проводились работы по созданию словарей и индексированию в целом на русском. Поэтому в качестве тестовой была взята коллекция, предоставленная МГУ для проектирования поисковых систем. Поскольку эксперименты проводились на тестовой коллекции с качественными ИР, показатели эффективности классификации несколько завышены. Но в любом случае классификатор считается удовлетворительным, если его F_1-мера превышает 0,6. Тестирование метода опорных векторов и его предложенной модификации проводилось на одной и той же коллекции, ИР проходили одинаковую предварительную обработку и оцифровку, что дает право на сравнение показателей эффективности методов классификации и приводит к выводу о некотором превосходстве метода опорных векторов на предложенном ядре матрицы взаимосвязи терминов во множестве ИР.

1. *Гриценко В.И.*, *Духновская К.К.*, *Урсатьев А.А*. Поисковый сервис. Проблемы, технологии, перспективы. // УСиМ. – 2006 – № 2. – С. 81–92.
2. *Винберг Э.Б.* Курс алгебры. – М.: Факториал Пресс, 2001. – 544 с.
3. *Википедия*. – http://ru.wikipedia.org/wiki/
4. *Хедли Дж.* Нелинейное и динамическое программирование. – М.: Мир, 1967. – 508 с.
5. *Platt J.C*. Fast training support vector machines using sequential minimal optimization // MIT Press – 1999. – P. 185–208.
6. *Карманов В.Г*. Математическое программирование. – М.: Наука, 1986. – 288 с.
7. *Fabrizio Sebastiani*. Machine Learning in Automated Text Categorization // ACM Comp. Surveys. – 2002. – **34**, – № 1. – P. 1–47.
8. *Лифшиц Ю*. Курс лекций «Алгоритмы для Интернета». – // http://yury.name/internet/

Поступила 11.10.2013
Тел. для справок: +38 067 194-7186 (Киев)
E-mail: *duchnov@ukr.net*