Теоретическая обоснованность применения метрических методов классификации с использованием динамического выравнивания (DTW) к пространственно-временным объектам*

***И. О. Автор*[1], *И. О. Соавтор*[2], *И. О. Фамилия*[1,2]**
author@site.ru; co-author@site.ru; co-author@site.ru
[1]Организация, адрес; [2]Организация, адрес

Решается задача выравнивания пространственно-временных рядов, строится функция расстояния. Для этого используется метод динамического выравнивания. В работе исследуется корректность метода динамического выравнивания и его модификаций к пространственно-временным рядам. При доказательстве, проверяют, что функция, создаваемая алгоритмом динамического выравнивания, является ядром. Проверка этого факта осуществляется при помощи теоремы Мерсера, основная часть которой заключается в проверке матрицы попарных расстояний на неотрицательную определенность. Также производится анализ зависимости качества классификации методом опорных векторов и методом k-ближайших соседей от различных функций.

Ключевые слова: *динамическое выравнивание; пространственно-временные ряды; ядро функции; теорема Мерсера; функция расстояния; функция расстояния; метод опорных векторов; метод k-ближайших соседей*

DOI: 10.21469/22233792

1 Введение

Функция расстояния между временными рядами может быть задана различными способами: Евклидово расстояние [5], метод динамического выравнивания временных рядов [3, 7], метод, основанный на нахождение наибольшей общей последовательности [10]. В [4] показано, что разность между значениями временного ряда, соответствующими различным временным отсчетам, не может рассматриваться в качестве описания расстояния между двумя объектами: эта метрика чувствительна к шуму и локальным временным сдвигам. Предлагается использовать метод динамического выравнивания временных рядов (англ. Dynamic Time Warping) [6]. Как показано в [9], этот метод находит наилучшее соответствие между двумя временными рядами, если они нелинейно деформированы друг относительно друга – растянуты, сжаты или смещены вдоль оси времени. Метод DTW используется для определения сходства между ними и введения расстояния между двумя временными рядами.

На данный момент существуют теоретические обоснования (создаваемая алгоритмом функция точно аппроксимирует аналитическую функцию) применения DTW лишь для некоторых временных объектов, например, для дизартрического распознавания речи с разреженными обучающими данными [11]. В этой статье теоретически обосновывается его применение для пространственно-временных объектов. Исследование проводится на данных [1].

*Работа выполнена при финансовой поддержке РФФИ, проекты № №00-00-00000 и 00-00-00001.

Алгоритм построения оптимальной разделяющей гиперплоскости – алгоритм линейной классификации [2]. В основе создания же нелинейного классификатора лежит замена скалярного произведения $\langle x, x' \rangle$ на функцию ядра $K(x, x')$. Таким образом осуществляется переход в спрямляющее пространство (kernel trick), который позволяет построить нелинейные разделители. Если изначально выборка была линейно неразделимой, то при удачном выборе ядра можно избавить от этой проблемы. Это позволяет применять линейные алгоритмы классификации (SVM) в случаях, когда выборка не разделяется линейно. Критерием функции ядра является теорема Мерсера [8],

2 Постановка задачи

В работе мы будем проверять выполнение условий теоремы Мерсера на разных данных для разных модификаций DTW. То есть, следующие два условия для функции $K(x, x')$, порожденной DTW:

- $K(x, x') = K(x', x)$
- $\int\limits_X \int\limits_X K(x, x') g(x) g(x') dx dx' \geqslant 0 \quad \forall g : X \to \mathbb{R}$

Последнее условие эквивалентно тому, что для любых наборов $\{x_1, \ldots, x_n\}$ матрица $K = ||K(x_i, x_j)||_{i,j}$ неотрицательно определена: $v^T K v \geqslant 0 \;\; \forall v \in \mathbb{R}^n$

В нашей задаче мы будем исследовать, насколько качественно функции (являющиеся ядрами или нет), полученные в результате DTW, подставленные в алгоритм SVM (Support Vector Machine), классифицируют объекты. Для начала рассмотрим задачу классификации объектов $X \in \mathbb{R}^n$ на два непересекающихся класса $Y = \{-1, +1\}$. Обучающая выборка $X^l = (x^j, y^j)_{j=1}^l$ Линейный классификатор будет иметь вид:

$$a(x) = sign(\sum_{i=1}^{n} w_i x_i - w_0) = sign(\langle w, x \rangle - w_0)$$

Тут использованы обозначения: $x = (x_1, \ldots, x_n)$ – признаковое описание объекта x, $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$ и $w_0 \in \mathbb{R}$ являются, так называемыми весами, являющимися параметрами алгоритма.

Заметим, что $\langle w, x \rangle = w_0$ задает в пространстве гиперплоскость, которая разделяет классы.

Заметим, что линейный классификатор $a(x)$ не изменится если w и w_0 умножить на одну и ту же положительную константу. Поэтому произведем нормировку наиболее удобным для нас способом: чтобы для всех ближайших к раздеяющих гиперплоскости объектов $x^j \in X^l \hookrightarrow \langle w, x^j \rangle - w_0 = y^j$. Таким образом получим для всех $x^j \in X^l$:

$$\langle w, x^j \rangle - w_0 \; \begin{cases} \leqslant -1, \; if \; y^j = -1 \\ \geqslant 1, \; if \; y^j = +1 \end{cases}$$

$-1 < \langle w, x \rangle - w_0 < 1$ задает полосу, которая разделяет классы. Внутри этой полосы нет ни одной точки обучающей выборки X^l, на ее границе лежат точки, ближайшие к разделяющей гиперплоскости. Сама гиперплоскость проходит ровно посередине полосы. Мы хотим добиться максимальной ширины h этой полосы (для более качественной классификации). Пусть точки x_-, x^+ – соответственно точки классов $-1, +1$, лежащие на границе полосы, тогда: $h = \frac{1}{||w||} \langle x_+ - x_-, w \rangle = \frac{\langle x_+, w \rangle - \langle x_-, w \rangle}{||w||} = \frac{(w_0 + 1) - (w_0 - 1)}{||w||} = \frac{2}{||w||}$

Таким образом, нам необходимо решить следующую задачу оптимизации:

$$\begin{cases} \frac{1}{2}\langle w, w\rangle \to min \\ (\langle w, x^j\rangle - w_0)y^j \geqslant 1 \quad i = 1, \ldots, l \end{cases}$$

Решаем данную задачу при помощи теоремы Каруши-Куна Таккера:

$$\begin{cases} L(w, w_0, \lambda) = \frac{1}{2}\langle w, w\rangle - \sum\limits_{j=1}^{l} \lambda_j \big((\langle w, x^j\rangle - w_0)y^j - 1\big) \to \min\limits_{w,w_0} \max\limits_{\lambda} \\ \lambda \geqslant 0 \\ \lambda_j(\langle w, x^j\rangle - w_0) = 0 \quad j = 1, \ldots, l \end{cases}$$

$\lambda = (\lambda_1, \ldots, \lambda_l)$ – вектор двойственных переменных

Необходимым условием седловой точки Лагранжиана L, является равенство нулю его градиента, отсюда получаем: $w = \sum\limits_{j=1}^{l} \lambda_j y^j x^j,\ \sum\limits_{j=1}^{l} \lambda_j y^j = 0 \qquad (1)$

Из первого следует, что вектор весов w является линейной комбинацией таких векторов обучающей выборки X^l, для которых соответствующее $\lambda_j \neq 0$. Согласно условию дополнящей нежесткости, так как $\lambda_j \neq 0$, исходные ограничения типа неравенств должны превратиться в равенства, значит, эти объекты (векторы) находятся на границе полосы. Вектора для которых $\lambda_j = 0$ не лежат на границе, и не участвуют в сумме, значит, фалгоритм не изменился бы, если их не было бы в выборке.

Объект $(x_j, y_j) \in X^l$, для которого $\lambda_j > 0$ и $\langle w, x^j\rangle - w_0 = y^j$ назовем опорным вектором (англ. support vector)

Подставим (1) обратно в выражение для Лагранжиана, получим:

$$\begin{cases} -L = -\sum\limits_{j=1}^{l} \lambda_j + \frac{1}{2}\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l} \lambda_i \lambda_j y^i y^j \langle x^i, x^j\rangle \to \min\limits_{\lambda} \\ \lambda \geqslant 0 \\ \sum\limits_{j=1}^{\lambda} {}_j y^j = 0 \end{cases} \qquad (2)$$

Полученная задача имеет единственное решение, поскольку целевая функция является квадратичным функционалом, имеющим неотрицательно определенную квадратичную форму, значит, является выпуклым, ограничения также выпуклы.

После решения задачи мы можем определить вектор w по формуле $w = \sum\limits_{j=1}^{l} \lambda_j y^j x^j$ и $w_0 = med\{\langle w, x^j\rangle - y^j :\ \lambda_j > 0,\ j = 1, \ldots, l\}$

Тогда получим: $a(x) = sign\big(\sum\limits_{j=1}^{l} \lambda_j y^j \langle x^j, x\rangle - w_0\big)$

Вопрос о решении двойственной задачи (2) все еще остается открытым.

Все эти рассуждения имеют место в случае, когда выборка линейно разделима, если же она не является таковой, то необходимо перейти в пространство большей размерности, где уже она будет линейно разделима. Этот переход будет осуществляться засчет функции ядра. Скалярные произведения $\langle x, x'\rangle$, таким образом, везде заменятся на значение функции ядра в соответствующих двух точках $K(x, x')$.

3 Эксперимент

Производится тестирование различных модификаций алгоритма DTW на различных данных [1], затем осуществляется проверка того, является ли полученная в результате работы алгоритма функция ядром (при помощи Теоремы Мерсера).

Как показывают эксперименты, полученная функция в точности ядром не является. Поэтому осуществляется попытка достроить результат DTW до ядра. Например, использовать RBF: $K(s_1, s_2 = exp(-\gamma \cdot \rho_{dtw}(s_1, s_2))))$

4 Заключение

Желательн о, чтобы этот раздел был, причём он не должен дословно повторять аннотацию. Обычно здесь отмечают, каких результатов удалось добиться, какие проблемы остались открытыми.

Литература

[1] http://www.timeseriesclassification.com/dataset.php.

[2] Kristin P. Bennett and Colin Campbell. Support vector machines: Hype or hallelujah? *SIGKDD Explorations*, 2(2):1–13, 2000.

[3] Clifford J. Berndt D. J. Fusing dynamic time warping to find patterns in time series. In *// Workshop on Knowledge Discovery in Databases*, pages 359–370, 1994.

[4] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn J. Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *PVLDB*, 1(2):1542–1552, 2008.

[5] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. In *ACM SIGMOD Conference on the Management of Data*, pages 419–429, Minneapolis, USA, 1994.

[6] Eamonn Keogh and M. Pazzani. Scaling up dynamic time warping to massive datasets. In *Proceedings 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 1–11, 1999.

[7] Eamonn J. Keogh and Chotirat (Ann) Ratanamahatana. Exact indexing of dynamic time warping. *Knowl. Inf. Syst*, 7(3):358–386, 2005.

[8] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Royal Soc. (A)*, 83(559):69–70, November 1909.

[9] Chan P. Fastdtw Salvador S. Toward accurate dynamic time warping in linear time and space. In *Workshop on Mining Temporal and Sequential Data*, page 11, 2004.

[10] Michail Vlachos, Dimitrios Gunopulos, and George Kollios. Discovering similar multidimensional trajectories. In Rakesh Agrawal 0001 and Klaus R. Dittrich, editors, *ICDE*, pages 673–684. IEEE Computer Society, 2002.

[11] Vincent Wan and James Carmichael. Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data. In *INTERSPEECH*, pages 3321–3324. ISCA, 2005.

Поступила в редакцию 01.01.2017