Классификация временных рядов*

Карасиков М. Е.
karasikov@phystech.edu
Московский физико-технический институт

Работа посвящена задаче классификации временных рядов. Исследуются способы решения задачи классификации, основанные на выделении из временных рядов признаков. Временной ряд рассматривается как последовательность некоторых повторяющихся независимых сегментов, каждый из которых порождает свое признаковое описание. Для решения задачи классификации в качестве описаний временных рядов предлагается использовать восстановленные распределения признаков их сегментов. Проведено экспериментальное сравнение на реальных данных качества классификации временных рядов предложенными алгоритмами и наиболее распространенным методом ближайшего соседа с DTW метрикой.

Последние изменения: 00:50, 10 апреля 2015 г.

Ключевые слова: *временные ряды, признаковая классификация, дискретное преобразование Фурье, дискретное вейвлет-преобразование.*

Time series feature-based classification*

Karasikov M. E.
Moscow Institute of Physics and Technology

Введение

Временные ряды являются результатом проведения любых повторяющихся во времени измерений. Обзор по методам и проблемам анализа временных рядов дается в [1, 2]. Одними из основных методов анализа временных рядов являются прогнозирование [3, 4, 5], обнаружение аномалий [6, 7, 8], сегментация [9, 10, 11], кластеризация [12, 13, 14] и классификация [15, 16, 17, 18]. Последние годы связаны с ростом интереса к данной области, проявляющимся в предложениях новых методов анализа временных рядов — метрик [19, 9, 20, 21, 22, 23], алгоритмов сегментации [24, 25, 26, 27], кластеризации [28, 29, 30], и др.

Временным рядом x будем называть конечную упорядоченную последовательность:

$$x = [x^{(1)}, \ldots, x^{(t)}].$$

В данной работе рассматривается задача классификации временных рядов, которая возникает во многих приложениях (медицинская диагностика по ЭКГ [31, 32] и ЭЭГ [33], классификация физической активности по данным с акселерометра [34, 35], верификация динамических подписей [36, 37] и т. д.) и заключается в определении неизвестных классов временных рядов рассматриваемого множества.

Формально задача классификации в общем виде может быть поставлена следующим образом. Пусть X — множество описаний объектов произвольной природы, Y — конечное

Научный руководитель: Стрижов В. В.

множество меток классов. Предполагается существование целевой функции — отображения $y : X \to Y$, значения которого известны только на объектах обучающей выборки

$$\mathfrak{D} = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset X \times Y.$$

Требуется построить алгоритм $a : X \to Y$ — отображение, приближающее целевую функцию y на множестве X. Задачей классификации временных рядов будем называть задачу классификации, в которой объектами классификации являются временные ряды.

Этап построения информативного пространства признаков, позволяющего добиться заданной точности классификации, является одним из важнейших этапов решения задачи классификации.

Одним из способов построения пространства признаков в задаче классификации временных рядов является задание функции расстояния [19, 9, 20, 21, 22, 23] между временными рядами, позволяющего в качестве признаков взять расстояния до опорных объектов. Данный метод чрезвычайно распространен в силу того, что позволяет свести исходную задачу классификации временных рядов к задаче выбора метрики — функции расстояния. При удачном выборе метрики дальнейшая классификация может происходить при помощи простейших метрических алгоритмов классификации, например, методом ближайшего соседа [38]. Данный подход так же зарекомендовал себя в self-training [18] и graph-based [39, 40] методах частичного обучения.

Другой способ состоит в извлечении из каждого временного ряда набора признаков — его информативного описания, позволяющего строить точные классификаторы с хорошей обобщающей способностью. При этом возникает задача выбора модели для описания временного ряда []. Признаками могут быть буквально произвольные функции $f : X \to \mathbb{R}^n$ исходных объектов. В работе [41] предлагается использовать в качестве признаков статистические функции (среднее, отклонения от среднего, коэффициенты эксцесса и др.). Стоит заметить, что необходимого качества классификации при таком подходе к построению пространства признаков часто удается добиться путем выбора соответствующих конкретной задаче признаков (см. пример [42]). В работе [43] в качестве признаков предлагается использовать коэффициенты дискретного преобразования Фурье (DFT). В [43, 44] предлагается использовать дискретное вейвлет-преобразование (DWT), которое сравнивается с предыдущими методами. Особенно эффективно DFT и DWT работают на квазипериодических рядах, где проявляется периодическая структура.

Под квазипериодичностью временного ряда будем понимать возможность выделения в нем характерных сегментов — периодов, то есть возможность представления каждого временного ряда $x = [x^{(1)}, \ldots, x^{(t)}]$ последовательностью в определенном смысле похожих его сегментов $s^{(1)}, \ldots, s^{(p)}$:

$$s^{(1)} = [x^{(1)}, \ldots, x^{(t_1)}], \ \ldots \ , s^{(k)} = [x^{(t_{k-1}+1)}, \ldots, x^{(t_k)}], \ \ldots \ , s^{(p)} = [x^{(t_{p-1}+1)}, \ldots, x^{(t)}].$$

В таком случае будем писать $x = (s^{(1)}, \ldots, s^{(p)})$.

В нашей работе предлагается алгоритм классификации квазипериодических временных рядов на основе распределения параметров модели, описывающей сегменты временных рядов. Предлагаемый подход к классификации квазипериодических временных рядов в общем виде изложен в разделе 1. В разделе 1 представлены эксперименты на реальных данных по сравнению предложенного подхода с подходом, предложенным в [34].

Постановка задачи

Дано множество временных рядов $X = \{x_1, \ldots, x_\ell\}$, состоящих из сегментов — элементов некоторого метрического пространства (S, d):

$$x_i = (s_i^{(1)}, \ldots, s_i^{(p)}) \in S^p, \quad i = 1, \ldots, \ell;$$

множество меток классов Y и обучающая выборка $\mathfrak{D} \subset X \times Y$. Задана модель порождения сегментов временного ряда:

$$g : \mathbb{R}^n \times S \to S,$$

где S — пространство сегментов временных рядов.

Каждому сегменту $s_i^{(k)}$ временного ряда x_i поставим в соответствие его вектор признаков

$$\mathbf{f}(s_i^{(k)}) = \arg\min_{\mathbf{w} \in \mathbf{R}^n} d(g(\mathbf{w}, s_i^{(k)}), s_i^{(k)})$$

Задано некоторое семейство алгоритмов классификации $A = \{a : \mathbb{R}^{n \times p} \to Y\}$ и функция потерь

$$\mathscr{L} : X \times Y \to \mathbb{R}.$$

Найти алгоритм классификации $a^* \in A$, доставляющий минимум функционалу качества $Q(a, \mathfrak{D}) \in \mathbb{R}, \quad a \in A$:

$$\begin{aligned} a^* &= \arg\min_{a \in A} Q(a, \mathfrak{D}) = \\ &= \arg\min_{a \in A} \frac{1}{|\mathfrak{D}|} \sum_{(x_i, y_i) \in \mathfrak{D}} \mathscr{L}\left(a\left(\mathbf{f}(s_i^{(1)}), \ldots, \mathbf{f}(s_i^{(p)})\right), y_i\right). \end{aligned}$$

Вычислительный эксперимент

Вычислительный эксперимент

Заключение

Литература

[1] *Esling, P.* Time-series data mining / P. Esling, C. Agon // *ACM Comput. Surv.* — 2012. — December. — Vol. 45, no. 1. — Pp. 12:1–12:34. http://doi.acm.org/10.1145/2379776.2379788.

[2] *Fu, T.-c.* A review on time series data mining / T.-c. Fu // *Engineering Applications of Artificial Intelligence.* — 2011. — Vol. 24, no. 1. — Pp. 164–181.

[3] *Weigend, A. S.* Time series prediction: forecasting the future and understanding the past / A. S. Weigend // *Santa Fe Institute Studies in the Sciences of Complexity.* — 1994.

[4] *Brockwell, P. J.* Time series: theory and methods / P. J. Brockwell, R. A. Davis. — Springer Science & Business Media, 2009.

[5] *Tsay, R. S.* Analysis of financial time series / R. S. Tsay. — John Wiley & Sons, 2005. — Vol. 543.

[6] *Weiss, G. M.* Mining with rarity: a unifying framework / G. M. Weiss // *ACM SIGKDD Explorations Newsletter.* — 2004. — Vol. 6, no. 1. — Pp. 7–19.

[7] *Chin, S. C.* Symbolic time series analysis for anomaly detection: a comparative evaluation / S. C. Chin, A. Ray, V. Rajagopalan // *Signal Processing.* — 2005. — Vol. 85, no. 9. — Pp. 1859–1868.

[8] *Yankov, D.* Disk aware discord discovery: finding unusual time series in terabyte sized datasets / D. Yankov, E. Keogh, U. Rebbapragada // *Knowledge and Information Systems.* — 2008. — Vol. 17, no. 2. — Pp. 241–262.

[9] Segmenting time series: A survey and novel approach / E. Keogh, S. Chu, D. Hart, M. Pazzani // *Data mining in time series databases.* — 2004. — Vol. 57. — Pp. 1–22.

[10] *Geurts, P.* Segment and combine approach for non-parametric time-series classification / P. Geurts, L. Wehenkel // Knowledge Discovery in Databases: PKDD 2005. — Springer, 2005. — Pp. 478–485.

[11] *Nunthanid, P.* Parameter-free motif discovery for time series data / P. Nunthanid, V. Niennattrakul, C. A. Ratanamahatana // Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2012 9th International Conference on / IEEE. — 2012. — Pp. 1–4.

[12] *Jain, A. K.* Data clustering: A review / A. K. Jain, M. N. Murty, P. J. Flynn // *ACM Comput. Surv.* — 1999. — September. — Vol. 31, no. 3. — Pp. 264–323. http://doi.acm.org/10.1145/331499.331504.

[13] *Liao, T. W.* Clustering of time series data—a survey / T. W. Liao // *Pattern Recognition.* — 2005. — Vol. 38, no. 11. — Pp. 1857 – 1874. http://www.sciencedirect.com/science/article/pii/S0031320305001305.

[14] *Zolhavarieh, S.* A review of subsequence time series clustering / S. Zolhavarieh, S. Aghabozorgi, Y. W. Teh // *The Scientific World Journal.* — 2014. — Vol. 2014.

[15] *Bakshi, B.* Representation of process trends—iv. induction of real-time patterns from operating data for diagnosis and supervisory control / B. Bakshi, G. Stephanopoulos // *Computers & Chemical Engineering.* — 1994. — Vol. 18, no. 4. — Pp. 303–332.

[16] *Geurts, P.* Pattern extraction for time series classification / P. Geurts // Principles of Data Mining and Knowledge Discovery. — Springer, 2001. — Pp. 115–127.

[17] Human activity recognition using smart phone embedded sensors: A linear dynamical systems method / W. Wang, H. Liu, L. Yu, F. Sun // Neural Networks (IJCNN), 2014 International Joint Conference on / IEEE. — 2014. — Pp. 1185–1190.

[18] *Wei, L.* Semi-supervised time series classification / L. Wei, E. Keogh // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '06. — New York, NY, USA: ACM, 2006. — Pp. 748–753. http://doi.acm.org/10.1145/1150402.1150498.

[19] Querying and mining of time series data: Experimental comparison of representations and distance measures / H. Ding, G. Trajcevski, P. Scheuermann et al. // *Proc. VLDB Endow.* — 2008. — August. — Vol. 1, no. 2. — Pp. 1542–1552. http://dx.doi.org/10.14778/1454159.1454226.

[20] *Berndt, D. J.* Using dynamic time warping to find patterns in time series. / D. J. Berndt, J. Clifford // KDD Workshop / Ed. by U. M. Fayyad, R. Uthurusamy. — AAAI Press, 1994. — Pp. 359–370. http://dblp.uni-trier.de/db/conf/kdd/kdd94.html#BerndtC94.

[21] A novel bit level time series representation with implication of similarity search and clustering / C. Ratanamahatana, E. Keogh, A. J. Bagnall, S. Lonardi // Advances in knowledge discovery and data mining. — Springer, 2005. — Pp. 771–777.

[22] *Salvador, S.* Toward accurate dynamic time warping in linear time and space / S. Salvador, P. Chan // *Intelligent Data Analysis.* — 2007. — Vol. 11, no. 5. — Pp. 561–580.

[23] *Marteau, P.-F.* Time warp edit distance with stiffness adjustment for time series matching / P.-F. Marteau // *Pattern Analysis and Machine Intelligence, IEEE Transactions on.* — 2009. — Vol. 31, no. 2. — Pp. 306–318.

[24] *Shatkay, H.* Approximate queries and representations for large data sequences / H. Shatkay, S. Zdonik // Data Engineering, 1996. Proceedings of the Twelfth International Conference on. — 1996. — Feb. — Pp. 536–545.

[25] *Li, C.-S.* Malm: a framework for mining sequence database at multiple abstraction levels / C.-S. Li, P. S. Yu, V. Castelli // Proceedings of the seventh international conference on Information and knowledge management / ACM. — 1998. — Pp. 267–272.

[26] *Vasko, K.* Estimating the number of segments in time series data using permutation tests / K. Vasko, H. Toivonen // Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. — 2002. — Pp. 466–473.

[27] Streaming time series summarization using user-defined amnesic functions / T. Palpanas, M. Vlachos, E. Keogh, D. Gunopulos // *Knowledge and Data Engineering, IEEE Transactions on.* — 2008. — July. — Vol. 20, no. 7. — Pp. 992–1006.

[28] *Fröhwirth-Schnatter, S.* Model-based clustering of multiple time series / S. Fröhwirth-Schnatter, S. Kaufmann // *Journal of Business & Economic Statistics.* — 2008. — Vol. 26, no. 1. — Pp. 78–89.

[29] *Corduas, M.* Time series clustering and classification by the autoregressive metric / M. Corduas, D. Piccolo // *Computational Statistics & Data Analysis.* — 2008. — Vol. 52, no. 4. — Pp. 1860 – 1872. http://www.sciencedirect.com/science/article/pii/S0167947307002368.

[30] *Cormode, G.* Conquering the divide: Continuous clustering of distributed data streams / G. Cormode, S. Muthukrishnan, W. Zhuang // Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. — 2007. — April. — Pp. 1036–1045.

[31] *Bortolan, G.* Diagnostic ecg classification based on neural networks / G. Bortolan, J. Willems // *Journal of electrocardiology.* — 1993. — Vol. 26 Suppl. — P. 75—79. http://europepmc.org/abstract/MED/8189152.

[32] Finding unusual medical time-series subsequences: Algorithms and applications / E. Keogh, J. Lin, A. W. Fu, H. Van Herle // *Information Technology in Biomedicine, IEEE Transactions on.* — 2006. — Vol. 10, no. 3. — Pp. 429–439.

[33] *Marcel, S.* Person authentication using brainwaves (eeg) and maximum a posteriori model adaptation / S. Marcel, J. d. R. Millán // *Pattern Analysis and Machine Intelligence, IEEE Transactions on.* — 2007. — Vol. 29, no. 4. — Pp. 743–752.

[34] Human activity recognition using smart phone embedded sensors: A linear dynamical systems method / W. Wang, H. Liu, L. Yu, F. Sun // Neural Networks (IJCNN), 2014 International Joint Conference on. — 2014. — July. — Pp. 1185–1190.

[35] *Kwapisz, J. R.* Activity recognition using cell phone accelerometers / J. R. Kwapisz, G. M. Weiss, S. A. Moore // *SIGKDD Explor. Newsl.* — 2011. — March. — Vol. 12, no. 2. — Pp. 74–82. http://doi.acm.org/10.1145/1964897.1964918.

[36] *Martens, R.* On-line signature verification by dynamic time-warping / R. Martens, L. Claesen // Pattern Recognition, 1996., Proceedings of the 13th International Conference on. — Vol. 3. — 1996. — Pp. 38–42 vol.3. http://dx.doi.org/10.1109/ICPR.1996.546791.

[37] *Gruber, C.* Signature verification with dynamic rbf networks and time series motifs / C. Gruber, M. Coduro, B. Sick // Tenth International Workshop on Frontiers in Handwriting Recognition / Suvisoft. — 2006.

[38] Pattern recognition and classification for multivariate time series / S. Spiegel, J. Gaebler, A. Lommatzsch et al. // Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data. — SensorKDD '11. — New York, NY, USA: ACM, 2011. — Pp. 34–42. http://doi.acm.org/10.1145/2003653.2003657.

[39] *Nguyen, M. N.* Positive unlabeled leaning for time series classification. / M. N. Nguyen, X.-L. Li, S.-K. Ng // IJCAI / Citeseer. — Vol. 11. — 2011. — Pp. 1421–1426.

[40] *Marussy, K.* Success: a new approach for semi-supervised classification of time-series / K. Marussy, K. Buza // Artificial Intelligence and Soft Computing / Springer. — 2013. — Pp. 437–447.

[41] *Nanopoulos, A.* Feature-based classification of time-series data / A. Nanopoulos, R. Alcock, Y. Manolopoulos // *International Journal of Computer Research.* — 2001. — Vol. 10. — Pp. 49–61.

[42] *Wiens, J.* Patient risk stratification for hospital-associated c. diff as a time-series classification task / J. Wiens, E. Horvitz, J. V. Guttag // Advances in Neural Information Processing Systems. — 2012. — Pp. 467–475.

[43] *Mörchen, F.* Time series feature extraction for data mining using dwt and dft. — 2003.

[44] *Zhang, H.* A non-parametric wavelet feature extractor for time series classification / H. Zhang, T. B. Ho, M. S. Lin // Advances in Knowledge Discovery and Data Mining. — Springer, 2004. — Pp. 595–603.