A Fast Time Series Classification through Sampling Approach

M. Arathi[1] and A. Govardhan[2]
[1]Associate Professor in Computer Science and Engineering, School of IT, JNTUH, Hyderabad, TS. INDIA. arathi.ms@jntuh.ac.in
[2]Professor, Department of Computer Science and Engineering, JNTUHCEH, Hyderabad, TS. INIDA. govardhan_cse@yahoo.co.in

Abstract

A time series data is set of values recorded at equal intervals of time. There are numerous areas where time series data is collected and need to be analyzed. One such analysis is classification of time series data. Given the labeled time series data, a model need to be generated which can classify the data. We witnessed many methods for time series data classification. One such method is base on shapelets. A time series subsequence is called shapelet if it has the most representative feature of a class. The shapelet based methods results in more accurate, more interpretable & fast classification. The problem with them is the high time complexity of training time. To overcome this, the sampling method is proposed. In this method, the most representative time series data objects of a class are picked up and the shapelets are extracted from these time series data. This greatly reduces the time complexity of training time. With experimental evaluation, it has been shown that the time complexity is significantly reduced as compared to existing methods.

1. Introduction

A time series data is sequence of values recorded at equal time intervals. In time series classification (TSC) process involves two steps. In first step, the labeled dataset is split into training and test sets. The classifier model is generated using training set. In second step, the model generated in first step is tested using test set [1]. The parameter usually used to assess the model is accuracy. In literature, there are many time series classification algorithms. Some algorithms consider whole sequence for classification and some consider a subsequence of time series. The later algorithms are more popular as they consume less time. One such algorithm is time series classification through shapelet. A shapelet is time series subsequence which contains the most representative features of a class. Since they do not consider the whole sequence for classification; they are more robust against noisy data. The training time of such algorithms is too high, even though the shapelets are generated offline [2]. This is because there are large number of shapelet candidates and the time complexity to assess each candidate is high. Almost all possible subsequences generated are selected as shapelet candidates [3, 4]. In order to decrease the training time of the classifier, Sampling through Splitting Approach (STSA) method has been proposed.

Unlike the other algorithms which try to reduce the shapelet discovery time [4, 5, 6, 7], the proposed algorithm tries to reduce the training data set through sampling approach, which in turn results in reduced number of shapelet candidates. Through experimental evaluation it has been shown that there is no reduction in accuracy of the classifier, though the training time has been greatly reduced.

The remainder of this paper is organized as follows. Section 2 gives some related works on TSC. The shapelet generation method is discussed in Section 3. The proposed STSA algorithm is introduced in Section 4. Experimental results are presented in Section 5 and our conclusions are given in Section 6.

2. Related Work

The time series classification through shapelets is introduced by Ye et. al. [8]. Since shapelets are interpretable, more accurate, and fast, they have quickly dominated the time series classification algorithms. Hence, there are many shapelet based algorithms. Some algorithms have worked on novel ideas for discovery of shapelets [7, 8, 9, 10]. This does not show up much improvement in accuracy. But they have less time complexity. Another class of algorithms uses shapelets to transform data. They have shown that transformation into this new data space improves the classification accuracy [3,4]. But they have very high time complexity. Yet another class of algorithms try to learn the shapelets in such a way that the accuracy of model is not effected and training time is reduced [11, 12, 13]. Our work is based on [8]. They have already included two pruning strategies. The first pruning strategy is Subsequence Distance Early Abandon. In this pruning strategy, once the distance computation between two time series exceeds the smallest distance computed so far, the computation is abandoned. It greatly reduces the runtime of algorithm. The second pruning strategy is called Admissible Entropy Pruning method. Here, it maintains the maximum gain obtained through some shapelet among the shapelet candidates examined so far. While examining the next shapelet candidate, it in the middle of the examining all the time series it goes for an optimistic approach and sees that whether this will improve the information gain of the shapelet. If it does so, then the remaining time series data distance computations are performed. Otherwise the shapelet candidate is pruned from further consideration. Through this pruning, the runtime is reduced by more than two orders of magnitude.

The time required to find the shapelet is reduced by reusing computations and pruning the search space[7].They cache the distance computations for future use, and then apply the triangle inequality to prune some candidates. Another work demonstrates that the time can be reduced by sharing the computation among different local shapelets in shapelet selection process [6]. Further work has been done in minimizing the time complexity. This was done by setting theminimumlengthofshapeletsto5 and if the length of time seriesisbiggerthan100, then the maximum length of the shapelets is set to half of the time series length. The step size is set to 3. In another work, the key points are used to find the shapelet candidates [14]. In [15], they also try to reduce the number of candidates by selecting only those subsequences which contained one or more important data points. Gordon et al. [16,17] introduce The SALSA-R algorithm (a shapelet sampling algorithm) for fast computation of shapelet-based classification trees which does not examine all possible shapelets. Renard et al. [18] proposed a random-shapelet (RS) to reduce dramatically the time required.

Their method is based on the randomization of the discovery process. Grabocka et al. [19, 20] process a Scalable shapelet Discovery (SD) which reduces the numbers of

shapletcandidatesthroughanonlineclusteringpruningtechnique.Karlssonetal.[21]reduced shapelet candidates by generalizing random shapelet forests (gRSF). However, these methods are aimed to speed up shapelet discovery algorithms.

3. Shapelet Generation

To generate the best shapelet that discriminates one class from another, we first need to generate the subsequences of all possible lengths for each time series data. Then each subsequence is examined to see if it has better discriminating powers. It uses information gain to evaluate each subsequence. Some computations can be pruned while computing distance between the time series T and subsequence S. Instead of computing the exact distance between every subsequence of T and the subsequence S, the computations can be ceased once the partial computation exceeds the minimum distance known so far. This is known as early abandon [22]. It also uses entropy pruning which reduces the time complexity of the algorithm.

It is possible to have different candidates same information gain. In such cases, either the longest candidate, the shortest candidate or the one that achieves the largest margin between the two classes can be selected as best candidate.

The classifier model used here is decision trees [23]. The nonleaf nodes of the decision tree contain shapelet and threshold value for splitting. The leaf nodes contain the class labels. At each nonleaf node we have binary split. The query subsequences are compared with the shapelet on nonleaf node and if the distance value is less than the threshold value then it goes to left subtree or else to the right subtree. This process continues till the leaf node is reached.

4. STSA Approach

In this paper, a Sampling Through Splitting Approach (STSA) has been introduced which tries to reduce the size of training data through sampling approach. Selecting one or some time series per class is insufficient and also reduces the accuracy of the model [24]. Hence the sample is selected by looking at the properties of each class.

Suppose there are n time series in one class $D = \{T_1, T_2, \dots, Tn\}$ and $Ti = \{t_1, t_2, , t_j, , t_m\}$ where m represents the length of time series. Firstly, a principle time series is selected. The principle time series is the time series which is closest to the mean sum value. The sum value of one time series is the sum of the individual values.

The sum value of all the time series in one class is added and then divided by n to obtain the mean value. The principle time series is the one which is closest to this mean value. After getting the criteria time series, Euclidean distance is computed between time series data in the class and principle time series. Then these distance values are sorted, and difference between the adjacent distance values are computed. Next, the standard deviation is computed for these values. At last, the data is split into groups by splitting at the sequence that has difference larger than half of the computed standard deviation. At last, one time series that has minimum sum distance is selected from every group.

5. Experiments and Evaluation

In this paper, 9 data sets from the UEA & UCR Time Series Classification Repository are selected [25]. Table 1 and Table 2 shows the results of execution of the algorithm. It is compared with other shapelet based classification algorithms such as Shapelet discovery algorithms (SD) [7, 8, 9, 10], Shapelet transformation algorithms (ST) [3, 4], Learning shapelets algorithms (LS) [16, 12, 13] and random-shapelet (RS) [18]. It is evident from Table 1 and Table 2 that the STSA approach is fastest among the methods (SD, LS, ST and RS) and does not result in reduction of the accuracy.

Table 1: Comparison of time (in seconds)

Dataset	SD	LS	ST	RS	**STSA**
coffee	4.12	5.23	4.124	3.49	**0.13**
Mallat	3.42	4.239	7.124	6.01	**1.34**
Gun	1.10	1.09	3.24	4.12	**0.24**
Swedish Leaves	3.21	2.16	3.498	2.078	**0.98**
ElectricalDevices	10.981	9.016	7.192	7.012	**4.39**
OSULeaves	2.143	3.18	2.109	1.905	**1.015**
ECGFivedays	0.429	0.761	0.298	0.391	**0.231**
ItalyPowerDemand	1.13	1.0876	1.90	2.00	**0.657**
Lighting7	6.80	6.178	4.956	3.96	**5.4**

Table 2. Misclassification rate of various shapelet based classification algorithms

Dataset	SD	LS	ST	RS	**STSA**
coffee	2.929	3.012	2.964	2.769	**2.01**
Mallat	7.981	8.801	8.021	7.991	**7.002**
Gun	19.321	18.90	18.395	18.513	**18.11**
Swedish Leaves	7.210	6.230	6.902	7.013	**3.01**
ElectricalDevices	16.908	16.709	15.994	15.856	**15.50**
OSULeaves	28.89	27.773	26.789	26.879	**26.78**
ECGFivedays	3.102	4.432	2.14	1.01	**0.126**
ItalyPowerDemand	9.910	9.999	8.992	8.987	**8.11**
Lighting7	4.987	4.442	3.99	3.997	**3.51**

6. Conclusion

Shapelet based time series classification has attracted a lot of researchers in time series data mining community. However, the time complexity of the shapelet selection process is too high due to large number of shapelet candidates generated and tested. In order to improve the training time, sampling approach STSA has been used with no accuracy reduced. In our algorithm, some time series data are selected from training data set and then find the best shapelet in this sample. This will lessen the number of shapelet candidates, which in turns reduces time complexity of algorithm. The results of experiments shows that STSA is faster than the original shapelet based classification algorithms with no accuracy reduced. Our results also demonstrate that our method is the faster method among the shapelet methods.

References

[1] Keogh, E., Kasetty, S., 2003. On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Mining and knowledge discovery 7, 349–371.

[2] Chang, K.W., Deka, B., Hwu, W.W., Roth, D., 2012. Efficient pattern-based time series classification on gpu, in: 2012 IEEE 12th International Conference on Data Mining (ICDM), IEEE. pp. 131–140.

[3] Lines, J., Davis, L.M., Hills, J., Bagnall, A., 2012. A shapelet transform for time series classification, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 289–297.

[4] Hills, J., Lines, J., Baranauskas, E., Mapp, J., Bagnall, A., 2014. Classification of time series by shapelet transformation. Data Mining and Knowledge Discovery 28, 851–881.

[5] Bagnall, A., Bostrom, A., Large, J., Lines, J., 2016a. The great time series classification bake off: an experimental evaluation of recently proposed algorithms. Extended Version.CoRR,abs/1602.01711.

[6] Xing, Z., Pei, J., Yu, P.S., Wang, K., 2011. Extracting interpretable features for early classification on time series, in: Proceedings of the 2011 SIAM International Conference on Data Mining, SIAM. pp. 247–258.

[7] Mueen, A., Keogh, E., Young, N., 2011. Logical-shapelets: an expressive primitive for time series classification, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 1154–1162.

[8] Ye, L., Keogh, E., 2009. Time series shapelets: a new primitive for data mining, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 947–956.

[9] Ye, L., Keogh, E., 2011. Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. Data mining and knowledge discovery 22, 149–182.

[10] Rakthanmanon, T., Keogh, E., 2013. Fast shapelets: A scalable algorithm for discovering time series shapelets, in: Proceedings of the 2013 SIAM International Conference on Data Mining, SIAM. pp. 668–676.

[11]Grabocka, J., Schilling, N., Wistuba, M., Schmidt-Thieme, L., 2014. Learning time-series shapelets, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 392–401.

[12] Shah, M., Grabocka, J., Schilling, N., Wistuba, M., Schmidt-Thieme, L., 2016. Learning dtw-shapelets for time-series classification, in: Proceedings of the 3rd IKDD Conference on Data Science, 2016, ACM.

[13] Kwok, L.H.J.T., Zurada, J.M., 2016. Efficient learning of time series shapelets.

[14] Zhang, Z., Zhang, H., Wen, Y., Yuan, X., 2016. Accelerating time series shapelets discovery with key points, in: Asia-Pacific Web Conference, Springer. pp. 330–342.

[15] Ji, C., Zhao, C., Pan, L., Liu, S., Yang, C., Wu, L., 2017. A fast shapelet discovery algorithm based on important data points. International Journal of Web Services Research (IJWSR) 14, 67–80.

[16] Gordon, D., Hendler, D., Rokach, L., 2012. Fast randomized model generation for shapelet-based time series classification. arXiv preprint arXiv:1209.5038.

[17] Gordon, D., Hendler, D., Rokach, L., 2015. Fast and space-efficient shapelets-based time-series classification. Intelligent Data Analysis 19, 953–981.

[18]Renard, X., Rifqi, M., Erray, W., Detyniecki, M., 2015. Random-shapelet: an algorithm for fast shapelet discovery, in: 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), IEEE. pp. 1–10.

[19] Grabocka, J., Wistuba, M., Schmidt-Thieme, L., 2015. Scalable discovery of time-series shapelets. arXiv preprint arXiv:1503.03238.

[20] Grabocka, J., Wistuba, M., Schmidt-Thieme, L., 2016. Fast classification of univariate and multivariate time series through shapelet discovery. Knowledge and Information Systems 49, 429–454.

[21] Karlsson, I., Papapetrou, P., Bostrom, H., 2016. Generalized random shapelet forests. Data Mining and Knowledge Discovery 30, 1053–1085.

[22] Keogh,E., Wei,L., Xi,X., Lee,S., and Vlachos, M., "LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures," *In the Proc of 32nd VLDB*, pp. 882-893, 2006.

[23] Breiman, L.,Friedman, J.,Olshen, R.A., and Stone, C.J., *Classification and regression trees*, Wadsworth, 1984..

[24] Sathianwiriyakhun, P., Janyalikit, T., Ratanamahatana, C.A., 2016. Fast and accurate template averaging for time series classification, in: 2016 8th International Conference on Knowledge and Smart Technology (KST), IEEE. pp. 49–54.

[25] The uea & ucr time series classification repository. URL www.timeseriesclassification.com.

AUTHORS

Mrs. M. Arathi pursued B.E.(CSE) from MVSREC, Hyderabad, Andhra Pradesh, India, in 2001 and M.Tech(CS) from JNTUH, Hyderabad, Andhra Pradesh, India, in 2008. Major field of study is data mining. She has worked as Assistant Professor in Sant Samarth Engineering College, Hyderabad, Andhra Pradesh from 2002 to 2003. Now she is working as Associate Professor in JNTUH, Hyderabad, Andhra Pradesh since 2003. She has 17 years of teaching experience. She has published papers in many national and international journals. She is a expert committee member in Institute for Innovations in Science and Technology. She has conducted a number of workshop and conferences on Data Science and Data Mining. She has been subject expert for Data Mining and Data Science. She has been judge for many paper presentation contests in JNTUH.

Prof. A. Govardhan pursued B.E.(CSE) from Osmania University, Hyderabad, Andhra Pradesh in 1992, M.Tech(CS) from JNU, New Delhi, India in 1994 and Ph.D(CS) from JNTU, Hyderabad, Andhra Pradesh in 2003. Areas of research include Databases, Data Mining and Information Retrieval Systems. He is presently a Director at SIT and Executive Council Member at Jawaharlal Nehru Technological University Hyderabad (JNTUH), India.

He has 2 Monographs and has guided 125 M.Tech projects, 20 Ph.D theses and has published 152 research papers at Journals/Conferences including IEEE, ACM, Springer, Elsevier and Inder Science. Delivered more than 50 Keynote addresses. He held several positions including Director of Evaluation, Principal, HOD and Students' Advisor. He is a Member on the Editorial Boards for Eight International Journals, Member of several Advisory & Academic Boards & Professional Bodies and a Committee Member for several International and National Conferences. He is a Chairman and Member on several Boards of Studies of various Universities and the Chairman of CSI Hyderabad Chapter. He is the recipient of 21 International and National Awards.