A Fast Time Series Classification using Shapelets

M. Arathi, and A. Govardhan

Abstract**—**Time series data are sequences of values measured over time. One of the most recent approaches to classification of time series data is to find shapelets within a data set. Time series shapelets are time series subsequences which represent a class. To classify time series data, the shapelets are identified which represent a particular class and discriminates it from other classes. Since one shapelet is not sufficient to classify entire data, two or more shapelets are identified which clearly distinguishes one class from other. The decision tree is used as classifier and each non leaf node of the decision tree contain shapelet and leaf nodes contain class label. To overcome the problem of overfitting the data in decision tree, we have used reduced error pruning. We show that by pruning the decision tree, a good increase in speed can be achieved without affecting the accuracy of classifier.

Keywords**—**Decision trees, Time Series Classification, Shapelets, Reduced Error Pruning.

I. Introduction

SINCE a decade there have been enormous papers on time series classification. One of the most promising recent approaches is to find shapelets within a data set [1]. The shapelets are time series subsequences which represent a particular class. Algorithms that are based on shapelets are interpretable, more accurate and significantly faster than state-of-the-art classifiers [2], [3].

There are two types of classification algorithms: algorithms that consider whole (single) time series sequence (global features) for classification and algorithms that consider a portion of single time series sequence (local features) for classification. Shapelets are local features of the time series data. In classification by shapelets, a shapelet that represents a particular class is identified. And then, instead of comparing the entire time series sequence, only small subsections of the two time series (shapelets) are compared. Because shapelets are small in size compared to original data, algorithms that use shapelets for classification, results in less time and space complexity. Shapelets have also been used successfully in many other applications, such as early classification [9], gesture recognition [10] and as a filter transformation for TSC [11].

For classification using shapelets, decision trees (binary) are used, where each nonleaf node represents a shapelet and leaf nodes represent class labels. To know how well the shapelet classifies the data, information gain [7] is used.

Apart from this, the other commonly used measures are such as the Wilcoxon signed-rank test [8], Kruskal-Wallis [12], Mood's Median [13] etc. The information gain/entropy measure is the better choice for two reasons. First, it can be easily generalized to the multiclass problem. Second, early entropy pruning can be done to avoid unnecessary distance calculations performed when finding the shapelet.

Before time series data are compared, they must be normalized to have mean as zero and a standard deviation of one [3]. It is meaningless to compare time series data with different offsets and amplitudes. The normalization of time series data can be performed by subtracting mean from each value of time series data and dividing the result by standard deviation of the data.

The sequence classification methods can be divided into three large categories.

- The first category is feature based classification, which transforms a sequence into a feature vector and then apply conventional classification methods. Feature selection plays an important role in this kind of methods.
- The second category is sequence distance based classification. The distance function which measures the similarity between sequences determines the quality of the classification significantly.
- The third category is model based classification, such as using hidden markov model (HMM) and other statistical models to classify sequences.

The rest of the paper is organized as follows. In Section II, we review related work. We propose a method in Section III. We report our experimental results in Section IV. We conclude our paper in Section V.

II. Related Work

A time series data is an ordered set of real-valued variables, where the data points are typically arranged by temporal order, spaced at equal time intervals.

The closest work is that of [1]. Here, the authors classify the time series data using shapelets. The first step in finding shapelets is to generate all possible subsequences of all possible lengths. A subsequence is part of the time series data having length less than or equal to the time series data. The minimum and maximum lengths for shapelets were computed using the simple cross-validation approach [24]. After generating all subsequences, each subsequence is tested to see how well it can classify the data. For this it generates an object histogram which contains all of the time series objects distances to the given subsequence. The histogram contains the values in increasing order of distance. To compute distance between two time series data, Euclidean distance measure is used. An optimization in

M. Arathi is with Jawaharlal Nehru Technological University Hyderabad, Hyderabad-500085, Andhra Pradesh, India. (E-mail: arathi.jntu@gmail.com).

A. Govardhan is with Jawaharlal Nehru Technological University Hyderabad, Hyderabad-500085, Andhra Pradesh, India. (E-mail: govardhan_cse@yahoo.co.in).

http://dx.doi.org/10.15242/IIE.E0414072

computing distance between the time series and subsequence is performed. That is, instead of computing the exact distance between every subsequence of a given time series data and the given subsequence, the distance calculations can be stopped once the partial computation exceeds the minimum distance known so far. This is known as early abandon [5]. If there is high probability of the subsequence resulting in best shapelet, then information gain is calculated. If the computed information gain is higher than best known so far information gain, then the subsequence is taken as best shapelet. The above process is repeated on all the subsequences.

To find information gain, the optimal split point for object histogram is computed. (An optimal split point is a distance threshold that has highest information gain as compared to other distance thresholds for given subsequence. The information gain is the difference between the entropy of dataset before splitting the data for a given split strategy and entropy of data after splitting the data.) Then the data is divided into two subsets by comparing the distance with optimal spit point. All the objects having distance less than split point are kept in one subset and the objects having distance greater than optimal split point are kept in other subset. And then information gain is computed.

Another optimization is performed to reduce the time complexity called entropy pruning. This is done during object histogram computation. One a distance value (computed between time series and the subsequence) is added to object histogram, it is checked to see if remaining calculations can be pruned. For this, the partially computed object histogram is taken. The remaining objects (for which the distance has not been computed to the given candidate) of one class are added to one end of the histogram and the objects of other class are added to the other end of the histogram and vice versa. Now, the information gain is computed. If it is greater than the best known so far information gain, then the histogram computation is continued, otherwise the remaining calculations with the subsequence are pruned.

It is often the case that different candidates will have the same best information gain. This is particularly true for small datasets. Such ties can be broken by favoring the longest candidate, the shortest candidate or the one that achieves the largest margin between the two classes.

Classifying with a shapelet and its corresponding split point produces a binary decision as to whether a time series belongs to a certain class or not. Because one shapelet is not sufficient to classify the entire time series data, a number of shapelets are used which clearly distinguishes one class from other. The shapelets are used along with distance threshold, which divides the data into two sets. The decision tree is used as classifier. The non-leaf nodes of the decision tree specify shapelet and distance threshold; and leaf nodes specify the class label. To find the accuracy of classifier, each time series data is fed into classifier, which moves it from root node to leaf node, which in turn gives the predicted class label. While moving from root to leaf node, the time series data is compared with every shapelet on the path using Euclidean distance measure. The predicted class label is compared with actual class label of the time series data. If they match, then count of number of correctly classified data is increased by one. Once all the data in test dataset are finished, the accuracy is computed as number of correctly classified data divided by total number of time series data in test dataset. To classify a time series data, it is fed into decision tree classifier, and the classifier returns the predicted class label.

Our focus is on to see the effect of decision tree pruning on speed of the algorithms. To the best of our knowledge, there is tremendous reduced in time complexity of classifying the data using decision tree classifier.

III. Proposed Method

We have built decision tree classifier for time series dataset using shapelets as explained Section II. We have observed that when a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of overfitting the data. Such methods typically use statistical measures to remove the least reliable branches. Pruned tree tend to be smaller and less complex and, thus, easier to comprehend. They are usually faster and better at correctly classifying independent test data than unpruned trees.

A. Decision Tree Pruning

In decision tree induction process, if we have tightly stopping criteria, it will lead to small and underfitted decision trees. On the other hand, if we have loosely stopping criteria, it will lead to generate large decision trees that are overfitted to the training set. Many pruning methods have been introduced to solve later problem [6]. There are two common approaches to tree pruning: prepruning and post pruning. In prepruning approach, a tree is pruned by halting its construction early. Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset tuples or the probability distribution of those tuples. In post pruning approach, some of the subtree are removed from a fully generated tree. A subtree at a given node is pruned by removing its branches and replacing it with a leaf. That is, given a decision tree classifier C and an inner (non-root, non-leaf) node t. Then pruning of C with respect to t is the deletion of all successor nodes of t in C which makes t a leaf node. The leaf is labeled with the most frequent class among the subtree being replaced. This process is repeated on all nonleaf nodes. The removal of the subtree should not result in reduction of the accuracy of the decision tree. Hence, it leads to a smaller and accurate decision tree. We have used reduced error pruning on the decision tree which is a post pruning method.

B. Reduced Error Pruning

A simple procedure for pruning decision trees, known as reduced error pruning, has been suggested by Quinlan [27]. While traversing over the internal nodes from the bottom to the top, the procedure checks for each internal node, whether replacing it with the most frequent class does not reduce the tree's accuracy. In this case, the node is pruned. The procedure continues until any further pruning would decrease the accuracy. We have used reduced error pruning on the generated decision tree classifier without scarifying accuracy of the tree. Hence, the algorithm generates a smallest accurate subtree.

IV. Experimental results

The experiments are conducted on UCR time series datasets such as wheat, mallet, coffee, gun, projectile points, historical documents, beef, car etc. [26]. Apart from optimizations like early abandon and entropy pruning, since we are also using decision tree pruning, our method has shown great reduction in time complexity for classifying the time series data using shapelets. On all the datasets, our proposed method has shown around 10 – 12% increase in speed.

The wheat dataset consists of 775 spectrographs of wheat samples grown in Canada between 1998 and 2005. There are different types of wheat, such as Soft White Spring, Canada Western Red Spring, Canada Western Red Winter, etc. The wheat dataset composes of all the above mentioned wheat types. The class label given for this problem is the year in which the wheat was grown. For this dataset, our method has shown 11% increase in the speed in testing phase and also in classification of unseen data.

There has been extensive study on Gun/NoGun motion capture time series dataset [2], [25]. This data has two classes. The classification algorithm should be able to identify whether the actor is holding gun or not. The difference between the two classes can be identified if we observe the time series data of the actor how he/she puts his/her hand down by his/her side. Our method has shown 9% increase in speed in testing phase and also in classification of unseen data for Gun/NoGun problem. Hence, the proposed method is faster than the existing method.

V. Conclusion and Future Scope

We have classified time series dataset using shapelets. The shapelets are time series subsequences and are highly representative of a class. Because one shapelet is not sufficient to classify the data, a number of shapelets are used which clearly distinguishes one class from other. Each shapelet is associated with a distance threshold, which divides the data into two sets. The decision tree is used as classifier. The non leaf nodes of the decision tree specify shapelet and distance threshold; and leaf nodes specify the class label. To classify a time series data, it is fed into decision tree classifier, which moves it from root node to leaf node, which in turn gives the predicted class label. While moving from root to leaf node, the time series data is compared with every shapelet on the path using dissimilarity/distance measure. And we have performed decision tree pruning using reduced error pruning method. The pruning method reduces the size of the decision tree which leads to reduction in time taken in testing phase and also in classification of unseen data. In future, we would like to use alternative distance measures and perform a comparative study on it. We also wish to check how the algorithm will perform on reduced representation of time series dataset. There is also scope to do signature verification using the proposed method.

References

[1] Lexiang Ye and Eamonn Keogh, "Time Series Shapelets: A New Primitive for Data Mining," KDD'09, June 29–July 1, 2009.

[2] Ding,H., Trajcevski, G., Scheuermann,P., Wang, X., and Keogh,E., "Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures," In Proc of the 34th VLDB, 2008, 1542–1552.

[3] Keogh,E. and Kasetty, S., "On the need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration," In Proc' of the 8th ACM SIGKDD, 2002,102-111.

[4] Mahalanobis, Prasanta Chandra, "On the generalised distance in statistics," Proceedings of the National Institute of Sciences of India **2** (1), 1936, 49–55.

[5] 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, 2006, 882-893.

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

[7] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Elsevier Publisher, Second Edition, 296-300.

[8] Wilcoxon,F., "Individual Comparisons by Ranking Methods," Biometrics, 1945,1, 80-83. http://dx.doi.org/10.2307/3001968

[9] P.Yu K. Wang Z. Xing, J. Pei, "Extracting interpretable features for early classification on time series," Proc. 11th SDM, 2011.

[10] B.Hartmann and N.Link, "Gesture recognition with inertial sensors and optimized DTW prototypes," Proc. IEEE SMC, 2010.

[11] J.Lines, L.Davis, J.Hills, and A.Bagnall, "A shapelet transform for time series classification," Tech. report, University of East anglia, UK, 2012.

[12] W.H.Kruskal, "A Nonparametric test for the several sample problem," The Annals of Mathematical Statistics 23(1952), no. 4, 525 – 540.

[13] A.M.F. Mood, Introduction to the theory of statistics, 1950.

[14] C.Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast Subsequence Matching in Time Series Databases," In SIGMOD Conference, 1994.

[15] Geurts, P., "Pattern Extraction for Time Series Classification," In Proc of the 5th PKDD, 2001, 115-127.

[16] E.J.Keogh and C.A.Ratanamahatana, "Exact indexing of dynamic time wraping," Knowl. Inf. Syst., 7(3), 2005. http://dx.doi.org/10.1007/s10115-004-0154-9

[17] D. Gunopulos, and G. Kollios, "Discovering similar multidimensional trajectories," In ICDE, 2002.

[18] L. Chen and R. T. Ng, "On the marriage of Lp-norms and edit distance," In VLDB, 2004.

[19] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," In Sigmod conference, 2005.

[20] Frentzos, K. Gratsias, and Y. Theodoridis, "Index-based most similar trajectory search," In ICDE, 2007.

[21] M. D. Morse and J. M. Patel, "An efficient and accurate method for evaluating time series similarity," In SIGMOD Conference, 2007.

[22] Y. Chen, M. A. Nascimento, B. C. Oosi and A. K. H. Tung, "SpADe: On Shape-based Pattern Detection in Streaming Time Series," In ICDE, 2007.

[23] J. Abflag, H. -P. Kriegel, P. Kröger, P. Kunath, A. Pryakhin, and M. Renz, "Similarity search on time series based on threshold queries," In EDBT, 2006.

[24] J. Lines, L. Davis, J. Hills and A. Bagnall, "A shapelet transform for time series classification," Tech. report, University of East Anglia, UK, 2012.

[25] Xi, X., Keogh, E., Shelton, C., Wei, L., and Ratanamahatana, C. A., "Fast Time Series Classification using Numerosity Reduction." In the Proc of the 23rd ICML. 2006, 1033-1040.

[26] Datasets : www.cs.ucr.edu/~eamonn/time_series_data/

[27] J.R.Quinlan, "Simplifying Decision Trees," International journal of man-machine studies, Elsevier, 1987.

Born in Hyderabad, Andhra Pradesh, India on 8th October 1979. Pursued B.E.(CSE) from MVSREC, Hyderabad, Andhra Pradesh, India, in 2001. Pursued M.Tech(CS), 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 for 11 months. Next, she is working as Assistant Professor in JNTUH, Hyderabad, Andhra Pradesh. It is more than 10 years since she has been with JNTUH, Hyderabad, Andhra Pradesh. She has 1 journal, 3 international and 1 national publication.

Mrs. Arathi is a expert committee member in Institute for Innovations in Science and Technology. She has been judge for many paper presentation contests in JNTUH. She has been subject expert for QTP testing tool.

B.E.(CSE) from Osmania University, Hyderabad, Andhra Pradesh in 1992. M.Tech(CS) from JNU, New Delhi, India in 1994. 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.

Prof.A.Govardhan 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

http://dx.doi.org/10.15242/IIE.E0414072