ACCURATE TIME SERIES CLASSIFICATION USING SHAPELETS

M. Arathi and A. Govardhan

School of Information Technology, JNT University, Hyderabad,
Andhra Pradesh, India.

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. In order to compare two time series sequences, existing work uses Euclidean distance measure. The problem with Euclidean distance is that it requires data to be standardized if scales differ. In this paper, we perform classification of time series data using time series shapelets and used Mahalanobis distance measure. The Mahalanobis distance is a descriptive statistic that provides a relative measure of a data point's distance (residual) from a common point. The Mahalanobis distance is used to identify and gauge similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant. We show that Mahalanobis distance results in more accuracy than Euclidean distance measure.

KEYWORDS

Decision trees, Information gain, Mahalanobis distance measure, Time series classification, Shapelets.

1. 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 a small subsection 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].

DOI : 10.5121/ijdkp.2014.4204

For classification with 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.

In classification of time series dataset using shapelets [1], Euclidean distance [14] has been used as similarity measure to compare two time series. There are some drawbacks of Euclidean distance measure. Firstly, it requires the time series data to be standardized, if scales differ. Secondly, it requires the two time series to be of same length. Thirdly, it does not take correlation of data items into consideration. To overcome some of the above problems, we have used Mahalanobis distance measure. It takes into account the correlations of the data items and is scale-invariant. In classification, the correlation among the dataset plays the key role. Hence, it is obvious that the accuracy will improved if Mahalanobis distance measure is used instead of Euclidean distance measure.

To compare two time series data, a distance measure that is metric should be used. A distance measure is said to be metric, if it satisfies following properties: 1) $d(p, q) \geq 0$ for all p and q and $d(p,q)=0$ only if $p = q$. (Positive definiteness), 2) $d(p, q) = d(q, p)$ for all p and q. (Symmetry), 3) $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r.(Triangle Inequality) where $d(p, q)$ is the distance (dissimilarity) between points (data objects) p and q. Both the distance measures (Euclidean and Mahalanobis) are metric.

Some of the other distance measures are Dynamic Time Warping (DTW) [15, 16], distance based on Longest Common Subsequence (LCSS) [17], Edit Distance with Real Penalty (ERP) [18], Edit Distance on Real sequence (EDR) [19], DISSIM [20], Sequence Weighted Alignment model (Swale) [21], Spatial Assembling Distance (SpADe) [22] and similarity search based on Threshold Queries (TQuEST) [23].

Before time series data are compared, they must be normalized to have mean as zero and a standard deviation of one [3]. Because, 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.

The rest of the paper is organized as follows. In Section II, we review related work. We define and compare distance measures in Section III. We report our experimental results in Section IV. We conclude our paper in Section V.

2. 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].

Each subsequence is tested to see how well it can classify the data. For this it generates a 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 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 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 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, then the histogram computation is continued, otherwise the remaining calculations with the candidate 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 time series data 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 predicts the class label.

Our focus is on to see the performance of Mahalanobis distance measure in time series classification using shapelets. To the best of our knowledge, our method gives more accurate classification of time series data than the existing method.

The following formatting rules must be followed strictly. This (.doc) document may be used as a template for papers prepared using Microsoft Word. Papers not conforming to these requirements may not be published in the conference proceedings.

3. PROPOSED METHOD

There are two numeric measures to compare two data objects: similarity & dissimilarity. The similarity measure tells about how alike two data objects are. It is higher when objects are more alike. It often falls in the range [0, 1]. And the dissimilarity measure specifies how different the two data objects are. It is lower when objects are more alike. Minimum dissimilarity is often zero. In this paper, dissimilarity/distance measure is used to compare two data objects. We show that using Mahalanobis Distance measure instead of Euclidean distance measure improves the accuracy of the algorithm.

3.1. Euclidean Distance Measure

In mathematics, the Euclidean distance or Euclidean metric is the ordinary distance between two points that one would measure with a ruler, and is given by the Pythagorean formula. By using this formula as distance, Euclidean space (or even any inner product space) becomes a metric space. It is defined as,

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) of data objects p and q.

The advantage of Euclidean distance measure is its simplicity in computation. But, when variables are on different measurement scales, standardization is necessary to balance the contributions of the variables in the computation of distance. The Euclidean distance computed on standardized variables is called the standardized Euclidean distance.

3.2. Mahalanobis Distance Measure

The Mahalanobis distance is a descriptive statistic that provides a relative measure of a data point's distance (residual) from a common point. It is a unitless measure introduced by P. C. Mahalanobis in 1936 [4]. The Mahalanobis distance is used to identify and gauge similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant.

Given a time series $x^{(k)}$, let the i^{th} data point be $x_i^{(k)}$. First, compute the(sample) covariance matrix C = (c_{ij}) of a family of time series $x^{(1)}$, $x^{(2)}$,…, $x^{(N)}$ of lengths n by $c_{ij} = \frac{1}{N-1}\Sigma_{k=1}^{N}(x_i^{(k)} - \bar{x}_i)(x_j^{(k)} - \bar{x}_j)$ where N is the number of instances and where $\bar{x}_i$ is the average of the i^{th} data point of the time series($(\bar{x}_i = \frac{1}{N}\Sigma_{k=1}^{N} x_i^{(k)})$.

The Mahalanobis distance measure is a special case of the generalized ellipsoid distance measure $D_M(x, y) = (x-y)^T M(x-y)$ where M is proportional to the inverse of the covariance matrix i.e., M α C^{-1}. Though the Mahalanobis distance measure is often defined by setting M to the inverse of the covariance matrix ($M = C^{-1}$), it is convenient to normalize it when possible so that the determinant of the matrix M is one: M = $(\det(c))^{\frac{1}{n}} c^{-1}$ where n is the length of the time series. The Mahalanobis distance measure minimizes the sum of distances between time series $\Sigma_{x,y} D_M(x, y)$ subject to a regularization constraint on the determinant (det(M) = 1). In this sense, it is optimal.

When the covariance is non-singular (det(C) ≠ 0) then the covariance is positive definite, and so is the matrix M: it follows that the square root of the generalized ellipsoid distance measure is a metric. That is, we have $D_M(x, y) = 0 \Leftrightarrow x = y$, it is symmetric, non-negative and it satisfies the triangle inequality $\sqrt{D_M(x, z)} + \sqrt{D_M(z, y)} \geq \sqrt{D_M(x, y)}$.

3.3. Euclidean vs. Mahalanobis distance measure

The Mahalanobis distance takes the co-variances into account, which lead to elliptic decision boundaries in the 2D case, as opposed to the circular boundary in the Euclidean case. The Euclidean distance may be seen as a special case of the Mahalanobis distance with equal variances of the variables.

The Mahalanobis distance is a fine way to reduce linear correlation and some scaling, so if one is looking at distance and has enough data, it makes more sense than Euclidean. In statistics, sometimes the nearness or farness is measured in terms of the scale of the data. Often scale means standard deviation. For univariate data, an observation that is one standard deviation away from the mean is closer to the mean than an observation that is three standard deviations away.

For many distributions, such as the normal distribution, this choice of scale also makes a statement about probability. Specifically, it is more likely to observe an observation that is about one standard deviation from the mean than an observation that is several standard deviations away. This is because the probability density function is higher near the mean and nearly zero as you move many standard deviations away.

For normally distributed data, the distance from the mean can be specified by computing the so-called z-score. For a value x, the z-score of x is the quantity $z = (x-\mu)/\sigma$, where μ is the time series data mean and σ is the standard deviation. This is a dimensionless quantity that can be interpreted as the number of standard deviations that x is from the mean.

The graph in Fig. 1 shows simulated bivariate normal data that is overlaid with prediction ellipses. The ellipses in the graph are the 10% (innermost), 20%, and so on till 90% (outermost) prediction ellipses for the bivariate normal distribution that generated the data. The prediction ellipses are contours of the bivariate normal density function. The probability density is high for ellipses near the origin, such as the 10% prediction ellipse. The density is low for ellipses are further away, such as the 90% prediction ellipse.

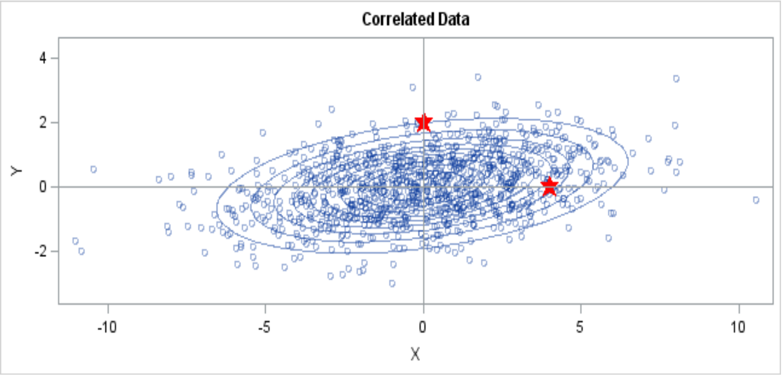

Fig. 1. Bivariate normal data with predicted ellipses.

In the graph, two observations are displayed by using red stars as markers. The first observation is at the coordinates (4,0), whereas the second is at (0,2). To see which mark is closer to origin, let us consider the two distance measures. The Euclidean distances are 4 and 2, respectively. Hence, according to Euclidean distance measure, the point at (0,2) is closer to the origin. However, for this distribution, the variance in the Y direction is less than the variance in the X direction, so in some sense the point (0,2) is more standard deviations away from the origin than the point (4,0).

Notice the position of the two observations relative to the ellipses. The point (0, 2) is located at the 90% prediction ellipse, whereas the point at (4,0) is located at about the 75% prediction ellipse. It means that the point at (4,0) is closer to the origin in the sense that you are more likely to observe an observation near (4,0) than to observe one near (0,2). The probability density is higher near (4,0) than it is near (0,2). Hence, according to Mahalanobis distance, the point at (4,0) is closer to origin than the point at (0,2).

In this sense, prediction ellipses are a multivariate generalization of units of standard deviation. The bivariate probability contours can be used to compare distances to the bivariate mean. A point p is closer than a point q if the contour that contains p is nested within the contour that contains q.

The Mahalanobis distance has the following properties: 1) It accounts for the fact that the variances in each direction are different. 2) It accounts for the covariance between variables. 3) It reduces to the familiar Euclidean distance for uncorrelated variables with unit variance.

4. Experimental Results

The experiments are conducted on standard datasets such as wheat, mallet, coffee, gun, projectile points, historical documents, beef, car etc. [26]. On all the datasets, our proposed method has shown around 10 – 15% increase in accuracy.

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 12% increase in the accuracy as shown in Fig. 2.

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 puts his hand down by his side. Our method has shown 8% increase in accuracy for Gun/NoGun problem as shown in Fig. 3. Hence, the proposed method has more accuracy than existing method.

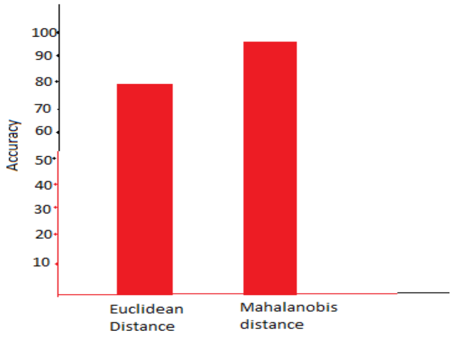

Fig. 2. Accuracy for wheat dataset using Euclidean vs Mahalanobis distance.

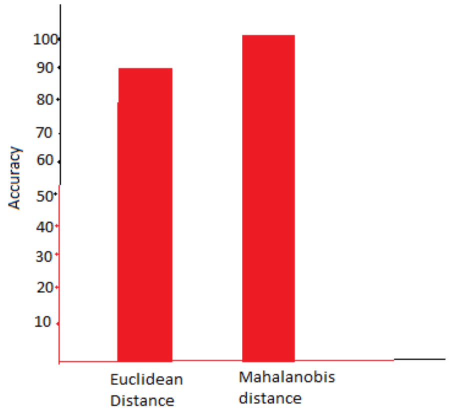

Fig. 3. Accuracy for Gun/NoGun dataset using Euclidean vs Mahalanobis distance.

5. 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. 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 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 Mahalanobis distance measure. Mahalanobis distance measure is a good choice for classification as it takes the correlation of data items into consideration and is scale in-variant. Hence, it is obvious that Mahalanobis distance measure will give more accurate results. We have also shown with experiments that the distance measure results in more accuracy than the Euclidean distance measure. In future, we would like to compare it with other distance measures. We are 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, (2009) "Time Series Shapelets: A New Primitive for Data Mining," KDD'09, June 29–July 1.

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

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

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

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

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

[7] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Elsvier Publisher, Second Edition.

[8] Wilcoxon,F., (1945) "Individual Comparisons by Ranking Methods," Biometrics, 1, 80-83.

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

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

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

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

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

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

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

[16] E.J.Keogh and C.A.Ratanamahatana, (2005) "Exact indexing of dynamic time wraping," Knowl. Inf. Syst., 7(3).

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

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

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

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

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

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

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

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

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

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

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 Assistant Professor in JNTUH, Hyderabad, Andhra Pradesh since 2003. She has 11 years of teaching experience. She has 1 journal, 3 international and 1 national publication. She is a expert committee member in Institute for Innovations in Science and Technology. 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.