ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

Survey on Classification of Time Series Data

[1] Penugonda Ravikumar, [2] Devanga Ampabathini Susmitha
[1]Assistant Professor [2] Under Graduate
IIIT - RKValley RGUKT – A.P.

Abstract— Classification of time series data is one of the popular fields nowadays. Many algorithms exist to get accurate and faster result which speeds up the computation. Nowadays E-Commerce sites are providing many user convenient options like recommendations for the products that they are purchasing which are playing key role in attracting the users to visit their sites. All these are happening based on the classification of data. Home automation is one of the evolving technologies in which data storage is important. It will be convenient if we are able to compare test data with limited trained data by preserving accuracy. Some data such as national security, personal information, bank details must be very confidential which has to be stored very securely. These all comes under time series data and data such as images, weather reports, speeches, and satellites data can be converted into time series data. Classification of normal data can be done by normal methods like nearest neighbor but to classify Time Series data by taking all these above features into consideration is difficult because time series data has to convert into structured format and it has to be sort and compare according to time Some existing algorithms are quite good in classifying Time Series data by satisfying some of these features. There are many algorithms which can classify data of different types quite accurately. But all algorithms might not satisfy all factors. So based on type of data we have to choose appropriate algorithm.

***Index Terms*— time-series, classification, distance measures .**

I. INTRODUCTION

Classification of time series data is one of the interesting fields. In recent years, interest in Research on time series data has been exploded. We can view time series as a sequence of values observed and measured at successive time instants spaced uniformly. Data exploded each day such as data of web sites, environment, digital media, stock market, health reports, data related to space can convert into time series data. Mostly time series related tasks will be queries on content, anomaly detection, predictions, classification and segmentation. Time series data will be able to relieve complexity of data. Classification of time series data is a bit difficult but it gives result more accurately compared with normal classification. In normal classification regulation of data is not compulsory but in time series classification regulation of data is essential. And in time series classification we have to consider many aspects to get precise results like privacy preserving, giving results in less time.

II. CLASSIFICATION APPROACHES

Time series classification can widely do by three approaches.
They are,
1. Distance-based classificati
2. Featured-based classification
3. Model-based classification

There are other classification techniques also like fuzzy and fast classification of time series data which are discussed in this survey.

1. Distance-based classification:

[1] Many algorithms run based on Distance-based classification such as nearest neighbor algorithm. Different distance measures have to be found to measure distances between data. The choice of distance measures plays a key role in the accuracy of the classification algorithm. Even though Euclidean distance is largely used distance measure it has some limitations such that two sequences must be in same length and it can calculate only linear data and we can find dissimilarity between series we use Lpnorm technique when p=2 it become Euclidean distance.

To overcome they used Dynamic time warping distance (DTW) measure to calculate non-linear data. However DTW should meet some local constraints such as boundary, monotonicity, continuity constraints. Classification of symbolic sequences such as DNA is also can be done by DTW. In this global alignment algorithm consider whole sequence where local alignment algorithms take sub-sequences and measures similarity. Needleman-Wunch algorithm is one of the Global alignment algorithms as it builds a matrix where each axis represents one of the two sequences and all values are initialized to zero. Now we can fill the matrix by applying below formula from bottom-right cell.

$$F(i, j) = \max \{ F(i-1, j-1)+s(xi, yj), F(i-1, j)-d, F(i,j-1)-d \}$$

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

Where s(x,y) is similarity of two characters in sequence and d is gap-open penalty.

[2]Though Euclidean distance is old and widely used distance measure for similarity search it has some limitations that it is much bricked it cannot measure similarity between two datasets if one of them is stretched or compressed. Even if we normalize data which is stretched or compressed Euclidean distance cannot measure similarity. This situation can be deal by Dynamic time warping distance (DTW). DTW is used for speech recognition. In this component shapes are same but they might not line up with axis. So to find similarity between such series we must wrap the time axis of one of the series. After DTW some algorithms presented which are optimal comparing with DTW. They are BLAST and FASTA. BLAST engine used for comparison of pairwise sequences. BLAST and FASTA can find result in optimal time but they cannot give guarantee to get optimal score. Sequential data can be multivariate so we can break them into separate series and get separate results by correlating those variables.

[2]Longest common subsequence (LCSS) similarity measure is one of the algorithms based on distance. It is used in text pattern matching and speech recognition. The main advantage of LCSS is, in Euclidean distance and DTW all elements must be utilized in calculation but in LCSS some data could be unmatched.

2. Feature-based classification:

[1]It is always not possible to work directly with the raw data that are highly noisy. Feature-based classification includes algorithms like decision-trees. This classification will be done based on feature-set. Sequential data has to transfer into feature-set before going classification algorithm. This transformation is done by Feature-based time series data classification techniques. Though most feature extraction methods are generic in nature, the extracted features are usually application dependent. That is, one set of features that work well on one application might not be relevant to another. Choosing appropriate features is challenging part in these techniques. Choosing correct features by manually gives more accurate results than by automatically. Patterns and wavelet decomposition are ways for extracting features from sequential data. Patterns should have local characteristics of a sequence and they are frequent in at least one class. Minimal Distinguishing Sub-sequence (MDS) algorithm is one of the pattern extraction algorithms. MDS allow for gaps with in the sub-sequences which makes classification more convenient. MDS converts shapes into sequential data to allow them to classify.

[2]Feature-extraction techniques where time-series data will be transferred into frequency domain are Discrete Fourier Transform(DFT), Discrete wavelet transform(DWT) and Singular Value Decomposition(SVD). In all these DWT is more commonly used technique since it stores time and frequency characteristics.

Frequency domain technique is one of the feature-extraction techniques data dimensionality can be reduced like in DWT. Kernal methods are also model for feature-extraction techniques that they deal with symbol sequences of different lengths and text data. DFT provides only frequency characteristics.

In [3] author proposed weighted feature-based classification is done by taking mean, max and min values of each class of trained data and find their score. They classified test data by comparing with trained data by considering weights. This is Simulated Annealing based approach finds the good approximation of the optimum weights assigned to the parameters. It combines different parameters which affect the classification and validation set is used to find the best values of the weights. After getting weights those weights are used to classify test dataset and to measure the classification accuracy.

3. Model-based classification:

[2]Time series is generated by some kind of model or by a mixture of underlying probability distributions. Time series are considered similar when the models characterizing individual series or the remaining residuals after fitting the model are similar.

[1]In model-based classification techniques, it constructs a model based on trained data and train the trained data on the model. Now compare and classify the test data based on that model that best fits it. Models can divide into many types. Models can be classified into statistical and neural network ones. Statistical models such as Gaussian, Markov models the probability distribution of the data. Artificial neural networks (ANN) are more related to statistical models. A recurrent neural network (RNN) is

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

special type of ANN, where there is a feedback connection in the network to keep track of its internal state when dealing with new inputs. Moreover RNN does not require knowledge of the data.
[2]Predictive models that tries to predict unavailable values of the data using the existing one, and descriptive models that tries to find patterns and relationships in the data models which are used a lot in sequence classification applications.

Fuzzy classification:
[4]The fuzzy rules discovered to handle fuzzy sets in order to handle the noises and fuzziness which plays a major role, while carrying out the classification on time series data. In this author proposed one more effective way of classification of data that is fuzzy algorithm. In this algorithm there are trained data which is already classified and test data which has to be classified based on trained data. In this procedure includes two steps Training phase and classification phase.
In training phase for every dimension of time series we calculate the min, max and mean of each class from training data which is used to perform the classification on test data.
Min ← which holds the min value of training data of particular class of particular dimension
Max ← which holds the max value of training data of particular class of particular dimension
Mean ← which holds the mean value of training data of particular class of particular dimension
In classification phase we measure the score for every class of the test data. In this technique to measure distance author used Euclidean and maxi-norm formulas. By score we can calculate score of each class. And finds the class to which the test pattern has the highest membership value and the test pattern is classified as belonging to the class.

Fast classification:
[5]In this author proposed fast classification of time series data which classifies data. As in fuzzy algorithm this algorithm also consists two phases that are training phase and classification phase. It finds out mean for each class of each dimension of trained data.
Mean ← which holds the mean value of training data of particular class of particular dimension
In classification phase they find out mean for each test case of each dimension and calculate distance with trained datasets means. In this method to calculate distance author used Euclidean and Maxi-norm formulas. For which trained dataset, test case achieved less distance test case will be assigned that class label. By this classification will be done.

III. CONCLUSION AND FUTURE WORK

In this paper, we have explained about the classification task using time series data. We have gone through different papers to get the know ledge about classification process, as a result, we have displayed our work in the form of a survey paper. We have observed several algorithms in different approaches like Distance based approach, Model based Approach, and Feature based approach. As per our observation Fast Classification time series algorithm is giving computationally accurate result when compared with the One Nearest Neighbor using Euclidean distance algorithm, which is a bench mark algorithm for the classification task. In order to reduce the computational time different authors have used different techniques by reducing the total comparisons of the test data while assigning a class label. In future, we would like to develop a novel classification algorithm using which, we have to get more accurate and much faster results, when compared with the standard algorithms like Nearest Neighbor.

REFERENCES

1. Tarek Amr "Survey on Time-Series Data Classification"
2. Kumar Vasimalla, Dept of Computer Science, Central University of Kerala "Survey on Time Series Data Mining"
3. Penugonda Ravikumar, Dept of Computer Science, RGUKT, "Weighted Feature-based Classification of Time Series Data"
4. Penugonda Ravikumar, Dept of CSA, IISC , "Fuzzy Classification of Time Series Data"
5. Penugonda Ravikumar, Dept of Computer Science, RGUKT, "Fast Classification of Time Series Data"