TIME SERIES CLASSIFICATION IN DATA MINING & CLUSTERING: A REVIEW

RAJASEKHAR NENNURI¹, M GEETHA YADAV², N SHALINI³

Department of CSE, IARE

ABSTRACT

various works in classification and clustering are done in recent times, we survey as an indication. In this paper we analyze the basics of time series mining, how time series mining is classified and general design data mining algorithms commonly used in time series data mining, and also explanation on the time series clustering algorithms. And also give the information of cluster based information.

1. INTRODUCTION

Values Taken by time by variable of time is known as time series. That means at a particular time interval to a specific time the values of specific data was taken that is known as a data of a time series of nᵗʰ. It is of two types,

- Discrete-time series
- Continues-time series

At a set of time intervals observations re produced then it is said to be Discrete time series [1]. If some time interval observations are read incessantly then it is Continues-time series.

2. TIME SERIES DATA MINING

Classification

As in classification [13], all the algorithms designed for assemblage of time-series knowing whichever commit to change the present this data algorithms to manage the written recorded data, or to improve the texture data itself for the existing algorithms. Undoubtedly dealing with time series data we noticed that new similarity measures are acceptable for the time series data. Whereas those doing adaptation on the sequence data either take out a feature-vector from it to be fed to the classifier, or start off with a dummy for the data. Limited their review to classification [6] algorithmic programs that think about providing new similarity measures, whereas on the opposite hand, categorized the classification algorithms into an analogous sorting to those of. Correspondingly, we have a tendency to area unit about to think about the classification algorithms within the accompanying order as “Distance-based classification”, “Feature-based classification”, “Model-based classification”

2.1 Distance based classification

It is classified based on distance between data components. Let’s take an example of K-nearest algorithm same as the distance between the data components are calculated. Based on the distance classification will be done. New parallel actions are required to apply conventional algorithm to time series data. Argues that distance measure choose the accuracy of classification algorithmic program. Sensitivity to distortion in time is accurate. Linear transformation is not sufficient because transformation is non-linear. Distance between two series is minimized by explaining Dynamic time wrapping[8] as a non-linear mathematical function. For this Dynamic time wrapping[8] which is an Elastic similarity measure is required to resolve this problem. By evaluating dynamic time wrapping dynamic programming dynamic time wrapping is calculate. Even though many investigators accepted the existence of DTW aloofness over Euclidean as its inefficiency in computation is adopting. Time complexity of DTW is (O(n^2)), this is will help in speeding up the calculations in DTW.
Hence their complexity is $O(n^2)$ [8] as on observation by more optimum algorithms such as BLAST and FASTA were developed by researchers. Heuristic approaches are used for newer algorithms, by this we can assume that by above algorithms there is no assurance of finding the score optimum. BLAST engine is used for BLAST 2.0 for pair wise serial contrast, is used when homologous series yet it is purposed as an option when comparing two series that are already known to be homologous. As noted earlier, sequential data can be multivariate. Noticed that breaking multivariate time series (MTS) [3] into separate series and processed each one on its own upshot in leaving out the relationship between those variable. They introduced a newer distance-measurement set of instructions, Eros in order to flock with MTS.

### 2.2 Feature-Based Classification

Feature based mostly classification algorithms, do their classification supported feature-set, example ANN and Decision Tree diagram. To use feature based mostly classification to time series data first we have to transform sequential data into the feature band. The choice of the suitable option is that the most arduous component [10] of this process, and at that post is all path a tradeoff between making out this procedure manually by domain experts are having it automated however less accurate in many shells. Patterns and wavelet rotting, as we are going to see these days for extracting features of sequential information.

Noticed that algorithmic rule that proves to identify tree-leaves supported their shapes are led astray by the distortion in their forms as result of insects eating parts of them. Rather than hoping
on the entire style of the leaves (global features), they chose native features (shapes) that significantly discriminates the leaves from totally different trees. They won over the shape information into a sequential one. The objective is to search out sub series, or shape lets as they west chadic them that are discriminating between classes. To see that subseries are to be taken, they ordered all series in keeping with their distance from all potential shape lets. And so they set out to seem for a midpoint that divides member series of every division. Having discriminative approach i.e. Binary decisions are taken total whether or not a replacement sequence belongs to a particular class or not, had to use a decision tree in their classifier. A lot of classes we have the more branches and split points have the tree. Similarly, introduced a pattern-extraction algorithmic rule called Minimal Distinguishing Subsequence (MDS). However, MDS permits for openings inside the sub-series, which hit it a lot of appropriate for classifying biological series as mentioned earlier.

Some other feature-extraction technique is to remodel the time-series information into the frequency domain, wherever the information spatial properties are often shortened. Listed DFT (Discrete Fourier Transform), DWT (Discrete Wavelet Transform) and SVD (Singular Value Decomposition) as examples here. Even so, researchers note that DWT is additional common in classification since it saves each time and frequency characteristics, whereas DFT provides the frequency characteristics solely. Such transformation additionally solves a drag mentioned within the first place, wherever we would like to review each local and broad trends among the sequence information.

DWT make great change the information into completely different frequency parts. The parts with higher order coefficients reflect the global trends of the data, whereas those with lower order coefficients reflect the local trends in it. Kernel methods (KM) are also just in feature extraction, in addition, they can amount with symbol-series with different measures end to end, was trading with wording facts as a personal bag of words rather than sequential data. Highlighted the power of core methods to amount with of, in the wording knowledge for computers without thought or attention of its very great number of features, ordinarily more than 10k. He was using Support Vector Machine in one, which is one of the core ways of doing. KM works out the inner product of the input vectors in a high dimensional space. By doing in this way, having an effect equal to the input decision boundaries can be made between the categories. Unlike, used KM to put in order wording as in sequential data. Like alignment-based distance measures, kernel methods are widely applied in biological series classification.
2.3 Model based classification

Example-based methods work by splitting the data into test data and breeding data, utilizing the training data construct a model and train the training dataset on the model to separate the training information. He split the models used in classification in statistical and neural network ones. Granting to the arithmetical models such as Gaussian, Poisson, Markov and Hidden Markov Models [4] [12], are manufactured. On the other hand, divided models into predictive models that attempts to predict unavailable values of the data using the existing one, and descriptive models that try to discover rules and relationships in the data, especially Markov models which are used a lot in sequence classification applications.

Hidden Markov Model (HMM) [12] is more successful in biological series classifications compared to Neural Networks, since it can deal with variable-length series, while the other technique requires fixed-length inputs. On the other hand, pinpointed some of HMM general limitations, criticize the assumption of states probability, independence, adding that HMM requires prior knowledge base-specific knowledge to take the input features. Broadly speaking, artificial neural networks (ANN) [5] are really close to statistical models. Defines recurrent neural networks (RNN) as a peculiar type of ANN, where there is a feedback connection in the web to maintain trail of its inner state when dealing with fresh inputs. RNN is suitable for sequential data since, according to, RNN is capable of modeling the temporal nature of the succession. Also, stated that in contrast to HMM, RNN [4] [5] does not need knowledge of the data. He too claimed that RNN is immune to temporal noise. However, as experienced in the first place, they require fixed-length inputs.

3. TIME SERIES CLUSTERING

Time series algorithms form cluster based along the type of time series information, distinctions can be established on whether the data are discrete-valued or real-valued, uniformly or non-uniformly sampled, univariate or variable, and whether data series are of equal or unequal length. Non-uniformly sampled data must be changed over into formally dressed data before clustering [3] operations can be done. Several algorithms have been built up to cluster different types of time series information. This paper groups previously developed time series clustering strategies into three non-worthy categories relying on whether they play directly with raw information, indirectly with features extracted from the crude data, or indirectly with models made from the new information.
• Raw-data-based clustering
• Feature-based clustering
• Model-based clustering

3.1 Raw-data-based clustering

Methods that work with sensitive information, either in the time or frequency recurrence area, are set into this class. The two time series being compared are normally tried out at the same interval, but their length might or might not be the same. For a clustering [2] multivariate [3] time varying data, modified the relocation clustering procedure that was originally prepared for stable information. To work a designated number of clusters, the best clustering among all the possible clustering [2] is the one with the minimum generalized Ward criterion function. A distance function based on the assumed independent Gaussian models of data inaccuracy and utilized a hierarchical clustering method to assemble seasonality series into a worthy turn of bunches. For the analytic thinking of dynamic biomedical image time series information, showed that deterministic annealing by the nominal free energy vector quantization (VQ) could be effectual. Proposed short time series (STS) [13] [5] distance to evaluate the similarity in shape formed by the proportional change of amplitude and the corresponding temporal information on uneven sampling intervals.

To group multivariate [3] vector series of earthquakes and mining explosions, applied hierarchical clustering as well as k-means clustering. The bundling of non stationary time series by applying local stationary renditions of K-L segregation information, measures that give optimal time–frequency statistics for measuring the deviation between two non-stationary time series. Modeled non-stationary time series with a time varying mixture of stationary sources, equivalent to the continuous hidden Markov model. A two-step procedure for clustering multivariate time series of equal or unequal length. The foremost step is applying the k-means or fuzzy, c-means clustering algorithm for time stripped data in order to convert multivariate real-valued time series into univariate discrete-valued time series. The second step employs the k-means or FCM algorithm again to group the converted univariate time series.

3.2 Feature-based clustering

It is always not possible to act instantly with the new data that are highly noisy. Several feature-based clustering methods have been suggested to address these fears. Though most feature extraction methods are non specific in nature, the extricated components are usually application
dependent. That is, one set of features that work considerably in single application might not be relevant to some other. Some studies even take another feature selection step to further cut back the number of feature dimensions after feature extraction. Modified the standard k-means clustering algorithm [7] for the identification of isolated words. To quantify the length between two spoken word patterns, a symmetric distance measure was determined based on the distance for measuring the space between two shapes. The projected modified k-means (MKM) clustering algorithm [8] was demonstrated to outperform the well established unsupervised without averaging (UWA) clustering algorithm at that fourth dimension. Cluster time series indirectly by applying two hierarchical clustering algorithms, the Ward’s minimum variance algorithm and the single linkage algorithm, to normalized spectra. The spectra were constructed from the original time series with the means adapted to zero. Clustered fMRI time series in gatherings of voxels with comparative enactments using two algorithms: k-means and Ward’s hierarchical clustering. The cross-correlation function between the fMRI activation and the prototype was utilized as the feature space described the utilization of self-sorting out maps for gathering data sequences segmented from the numerical time series using a continuous sliding window with the object to discover similar temporal patterns dispersed along the time series.

3.3 Model-based clustering

This category of approaches, considers that each time series is brought forth by some kind of model or by a mix of the underlying probability distributions. Time series are considered similar at this point when the models characterizing single arrangement or the rest of the residuals subsequent to fitting the model are comparative.

For clustering or choosing from a set of dynamic structures introduced the Euclidean distance [7] between their equivalent autoregressive extensions as the metric. Evaluated three meta-heuristic methods for partitioning a circle of time series into clusters. Prompted by questions raised in the context of musical performance theory, and defined hierarchical smoothing models to see the relationship between the emblematic structure of a music score and its operation, with each represented by a time series. Developed an agglomerative hierarchical clustering procedure that is grounded on the p-value of a test of hypothesis applied to every pair of given stationary time series, presented BCD: a Bayesian algorithm for clustering by dynamics. Studied the clustering of ARIMA time-series, by using the Euclidean distance amid the Linear Predictive Coding cepstra of two time series as their dissimilarity measure proposed a model-based process for clustering univariate ARIMA series. Taking on the Gaussian mixture model for speaker verification, anticipated a fuzzy, c means clustering-based normalization method to
obtain a better score to be compared with a given threshold for taking or refusing a claimed speaker.

4. CONCLUSION

We have gone over some major comes in time-series data processing. Since time-series information square measure sometimes terribly giant, discovering acquaintance of that immense information becomes a challenge, which works to monumental analysis challenges. The similarity live is incredibly vital a part of statistic data processing, that decides/figures out the accuracy of information mining task. We have a tendency to reassess a number of the vital works of your time series classification and agglomeration. we'd prefer to stress that the key live in any in data processing endeavor invariably lies in opting the proper illustration of information and similarity live for the project at hand.

References

1. Tak-chung Fu ,A review on time series data mining, Engineering Applications of Artificial Intelligence-elsevier, 2011
3. mohammed waleed kadous, Classification of Multivariate Time Series and Structured Data Using Constructive Induction, Springer, 2005
4. V.Kavitha, Clustering Time Series Data Stream – A Literature Survey- IJCSIS, 2010

