ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

AHAC: Decision Tree Classification with Agglomerative Hierarchical Algorithm Clustering for Time Series Data Clustering

S. Nithya[1], Dr. R.Gunavathi[2]

M.Phil Research Scholar, Dept. of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi, Tamil Nadu, India[1]

Associate Professor& Head, PG Dept. of Computer Applications, Sree Saraswathi Thyagaraja College, Pollachi, Tamil Nadu, India[2]

ABSTRACT: Clustering is the application of data mining techniques to discover patterns from the database. This research work incorporates soft clustering concept, which is the process of deriving the information from the similarity from the unsupervised database. The choice of a suitable distance measure is crucial to the clustering process and, given the vast number of distance measures for time series available in the literature and their diverse characteristics, this selection is not straightforward. To overcome these issues, this research work presents an optimal perspective on the problem of Similarity Measure Selection for Clustering Time Series Databases applications. The proposed method called "Agglomerative Hierarchical algorithm Clustering deviations for tree classification in large database", which takes as key measures of correspondence between pairs of data points. The proposed method is to establish a unified framework on unsupervised data sets.

KEYWORDS: Clustering;Time Series database; Hierarchical clustering;K-means.

I. INTRODUCTION

Data Mining, "The Extraction of hidden predictive information from large databases", is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Clustering is an unsupervised pattern recognition technique which automatically seeks to gather objects in "natural" groups. In addition to providing a natural classification, clustering gives an insight into the underlying structure of the data. The goal is that the objects in each cluster are similar while the different clusters are dissimilar. To do this we need to define a proximity measure that quantifies what we mean by these terms. A proximity measure is either a similarity measure or a dissimilarity measure, where a much used example of the latter is the Euclidean distance. The different ways of defining proximity is part of the reason why many different clustering procedures have been developed, ranging from simple heuristics suitable for a particular type of dataset to general iterative schemes which seeks to optimize some associated optimality criterion. Ideally one should use a clustering approach that produce good results in a wide variety of situations, since the general assumption for clustering is that we know little or nothing about the data in advance.

When clustering in the presence of time series data [1], the clustering method must be carefully selected based on the characteristics of the time series of primary interest. Consider the two time series,both of the time series are autoregressive of order two. A second order autoregressive process is characterized by relating an observation at time point *t*, say x_t, to the observations at the previous two time points, xt-1 and xt-2, with some additional error, typically

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

taken to be from a zero mean Gaussian process.If the process is of primary concern, then yes, the two time series should be clustered together. Both time series are generated with the same parameter set and both oscillate about their mean in a similar fashion. If employing a clustering algorithm that makes use of a dissimilarity matrix, the choice of discrepancy measure should focus on differences in the processes, or the underlying mechanisms, that produced the time series.

II. RELATED WORK

In [2] authors proposed a time-series data mining is to try to extract all meaningful knowledge from the shape of data. Even if humans have a natural capacity to perform these tasks, it remains a complex problem for computers. In this article authors intended to provide a survey of the techniques applied for time-series data mining. In [3] authors proposed a method for clustering of time series based on their structural characteristics. Unlike other alternatives, this method does not cluster point values using a distance metric, rather it clusters based on global features extracted from the time series. In [4] authors illustrated the boxplot is a very popular graphical tool to visualize the distribution of continuous uni-modal data. It shows information about the location, spread, skewness as well as the tails of the data. However, when the data are skewed, usually many points exceed the whiskers and are often erroneously declared as outliers. In [5] authors showed that binary relevance-based methods have much to offer, especially in terms of scalability to large datasets. The authors exemplified this with a novel chaining method that can model label correlations while maintaining acceptable computational complexity.In [6] authors to provided a timely review on this area with emphasis on state-of-the-art multi-label learning algorithms. Firstly, fundamentals on multi-label learning including formal definition and evaluation metrics are given. Secondly and primarily, eight representative multi-label learning algorithms are scrutinized under common notations with relevant analyses and discussions. In [7] authors considered the time series are ubiquitous, and a measure to assess their similarity is a core part of many computational systems. In particular, the similarity measure is the most essential ingredient of time series clustering and classification systems. Because of this importance, countless approaches to estimate time series similarity have been proposed. However, there is a lack of comparative studies using empirical, rigorous, quantitative, and large-scale assessment strategies.

III. PROPOSED ALGORITHM

The proposed algorithm accepts the data classification parameters as input which contains the MATLAB simulation where the novel Similarity Measure Selection for Clustering Time Series Databases algorithm is applied to the Synthetic and UCR dataset. This overall architecture in figure 1.1 follows automatic distance selection for time series databases from the start to end state.

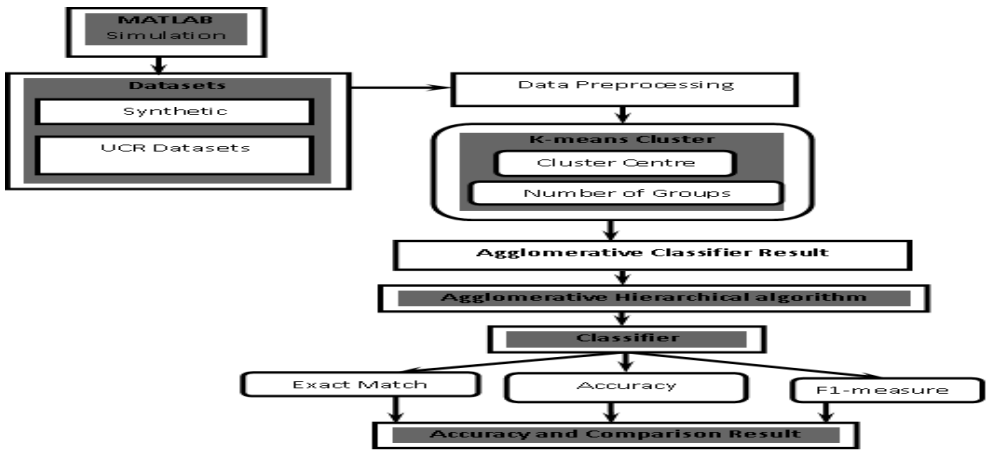


Fig.1. Architecture of Proposed System

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

A. *DATA PREPROCESSING*

Data pre-processing method is kind of data cleaning technique it plays a very important role in data classification techniques and applications. It is the first step in the adaptive relevance feature discovery mining process. In this process there are three key steps of procedures namely, Training set extraction, Feature Attribute selection and Filtering methods.

The data pre-processing of untrained raw dataset is first partitioned into three groups: (1) a predetermined set of instance initiation, (2) the group of attributes (features, variables) and (3) the class of attribute. For each groups in the dataset, a reduction decision classification is constructed. For each reduction system is consequently divided into two parts: the training dataset and the testing dataset. Each training dataset uses the corresponding input features and fall into two classes: normal (+1) and abnormal (−1).

The training set feature set process to compute the cross validation classification error for a large number of features and find a relatively stable range of small error. The feature range is denoted by Ω. The optimal number of features (denoted as n*) of the training set is determined with in Ω. The complete process includes three steps:

- The Training feature selection is to select n (a preset large number) sequential features from the input X. This leads to n sequential feature sets F1⊂F2⊂ … ⊂Fn-1⊂Fn.
- The n sequential feature sets F1, …, Fk, …, Fn,(1≤ k ≤n) to find the range of k, called Ω, within which the respective (cross-validation classification) error ek is consistently small (i.e., has both small mean and small variance).
- Within Ω, find the smallest classification error ek = min ek. The optimal size of the candidate feature set, n*, is chosen as the smallest k that corresponds to e*.

B. *K-MEANS CLUSTER TREE*

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume *k* clusters) fixed a priori. The main idea is to define *k* centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point re-calculate the *k* new centroids as centers of the clusters resulting from the previous step. After these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop *k* centroids change their location step by step until no more changes are done.

C. *AGGLOMERATIVE CLUSTER GROUP FORMATION*

The Agglomerative cluster group formation consists in training different classifiers with bootstrapped replicas of the original training data-set. That is, a new data-set is formed to train each classifier by randomly drawing (with replacement) instances from the original data-set (usually, maintaining the original data-set size).The bagging ensembles to deal with class imbalance problems due to its simplicity and good generalization ability. The hybridization of bagging and data pre-processing techniques is usually simpler than their integration in boosting. A bagging algorithm does not require re-computing any kind of weights; therefore, neither is necessary to adapt the weight update formula nor to change computations in the algorithm. In these methods, the key factor is the way to collect each bootstrap replica, that is, how the class imbalance problem is dealt to obtain a useful classifier in each iteration without forgetting the importance of the diversity.

D. *AGGLOMERATIVE HIERARCHICAL ALGORITHM CLUSTERING (AHAC)*

Agglomerative Hierarchical algorithm Clustering is not a new task and we had already the same concept in classical data clustering. The only difference is the difference between the nature of input values. In classical data as the feature values which can be presented in multi-dimensional vectors and therefore {x, y} represents two distinctive values in a 2-d space. With this definition spatial clustering can be simplified as a vector with two values like x y but this time

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

instead of values x and y, the longitude and latitude a object can be replaced. With this assumption the spatial clustering problem is exactly like clustering of 2-d vectors.

IV. PSEUDO CODE

ALGORITHMI: AHAC CLUSTERING

Let X = {x1, x2, x3, ..., xn} be the set of data points. AHAC requires two parameters: ε (eps) and the minimum number of points required to form a cluster (minPts).

Step 1: Start with an arbitrary starting point that has not been visited.
Step 2: Extract the neighbourhood of this point using ε (All points which are within the εdistance are neighbourhood).
Step 3: If there are sufficient neighbourhood around this point then clustering process startsand point is marked as visited else this point is labelled as noise (Later this point can become the part of the cluster).
Step 4: If a point is found to be a part of the cluster then its ε neighbourhood is also the partof the cluster and the above procedure from step 2 is repeated for all ε neighbourhood points. This is repeated until all points in the cluster is determined.
Step 5:A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.
Step 6: This process continues until all points are marked as visited.

V. SIMULATION RESULTS

The research work performed the experimentation, it is important to have some information about the characteristics of the databases that have been used in the study. The goal is to analyze the features and the labeling of the databases and to find the similarities and differences that could exist between the synthetic databases and thedatabases from the UCR. To evaluated the performance of all these classifiers using three metrics (Exact Match, Accuracy and F1-measure). The first is the Exact Match (EM) metric, which measures the proportion of correctly classified instances:

$$Exact\ Match\ (EM) = \frac{1}{N_{Te}} \sum_{i=1}^{N_{Te}} I\left(\hat{L}_i = L_i\right) \quad eqn.(1)$$

whereN_{Te} is the total number of instances in the testing set, L_i is the true set of class labels for instance i and $\hat{L}_i$ is the predicted set of labels for this same instance. Itakes a valueof 1 if the condition is true and 0 otherwise. It must be noted that EM is the strictest among the evaluation metrics in the multi-label framework because it only considers the instances whose predicted label set is identical to the ground truth label set. The Table 1 represents the Exact Match (EM) compare the proposed Density spatial clustering with existing approaches of Ensemble Classifier Chain (ECC) method [13], Random-*k*-labelsets classifier (RkL) [14], which transforms the multi-label problem into a multiclass framework.

Table 1: Exact Match (EM) values for the Multi-Label Classifiers

Groups	ECC	RkL	AHAC
Experiment 1	0.45	0.46	0.84
Experiment 2	0.42	0.44	0.75
Experiment 3	0.28	0.28	0.52
Experiment 4	0.07	0.13	0.65
Experiment 5	0.23	0.23	0.59

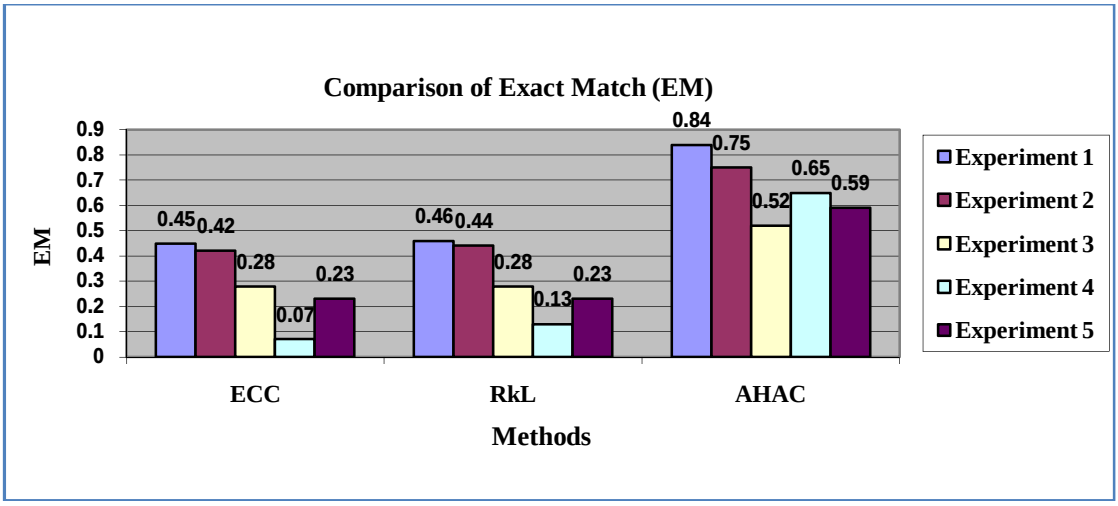
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016



VI. CONCLUSION AND FUTURE WORK

In this paper presents an enhanced method such as Similarity Measure Selection for Clustering Time Series Databases using Agglomerative Hierarchical algorithm Clustering (AHAC) based logical cluster tree method to solve the problem of Similarity Measure Selection high dimensional data clustering. By applying the Boolean class regression approach focuses on clustering time series databases and proactively exposes behavioral deviations by checking inside the class labels outputs. The proposed research work presents a new approach of Decision tree classification with Agglomerative Hierarchical algorithm Clustering (AHAC) framework automatically select the most suitable distance measures for clustering a time series database. To validate the proposed approach in a large scale UCR (Uniform Crime Reporting) bio-medical (ECG) datasets case study, and the results show that clustering approach can indeed serve as an accurate strategy for grouping regression deviations. A further challenge is to identify an important future direction is to develop a computationally efficient method of determining the distance metric of the embedding space, Manifold Finding and Dynamic/Streaming data.

REFERENCES

1. T. W. Liao, “Clustering of time series data: A survey,” Pattern Recog., vol. 38, no. 11, pp. 1857–1874, Nov. 2005.
2. P. Esling and C. Agon, “Time-series data mining,” ACM Comput. Surveys, vol. 45, no. 1, pp. 1–34, Nov. 2012.
3. X. Wang, K. Smith, and R. Hyndman, “Characteristic-based clustering for time series data,” Data Mining Knowl. Discovery, vol. 13, no. 3, pp. 335–364, May 2006
4. M. Hubert and E. Vandervieren, “An adjusted boxplot for skewed distributions,” Comput. Statist. Data Anal., vol. 52, no. 12, pp. 5186–5201, Aug. 2008.
5. J. Read, B. Pfahringer, G. Holmes, and E. Frank, “Classifier chains for multi-label classification,” Mach. Learning, vol. 85, pp. 333–359, 2011.
6. M.-L. Zhang and Z.-H. Zhou, “A review on multi-label learning algorithms,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
7. J. Serra and J. L. Arcos, “An empirical evaluation of similarity measures for time series classification,” Knowl.-Based Syst., vol. 67, pp. 305–314, Sep. 2014.