WEIGHTED TEMPORAL PATTERN MINING WITH DIMENSIONALITY REDUCTION USING MODIFIED AFCM TECHNIQUE

J. Mercy Geraldine¹ E. Kirubakaran² S. Sathiya Devi³

¹Research Scholar, Anna University, Chennai, Tamilnadu, India.
²SSTP(Systems), Bharat Heavy Electricals Ltd., Tiruchirappalli, Tamilnadu, India.
³Dept. of CSE, Bharathidasan Institute of Technology, Anna University, Tiruchirappalli, Tamilnadu, India.

ABSTRACT

Frequent itemset mining from a time series database is a difficult task. Various techniques have been proposed to mine the frequent associations among the data from the temporal database, but the huge size of the database and frequent time based updates to the database lead to inefficient frequent itemsets. Hence we proposed a dimensionality reduction method which reduces the quantity of data considered for mining. In the proposed system, initially the time based data are converted into fuzzy data. These fuzzy data are provided as input to the proposed Modified Adaptive Fuzzy C Means (MoAFCM) algorithm which is a combination of FCM clustering algorithm and Cuckoo search optimization algorithm. FCM performs dimensionality reduction on the fuzzy data and clustering is performed by the combination of both FCM and cuckoo search optimization algorithm leading to optimized clusters. The resulting clusters contain reference points instead of the original data. Optimization by cuckoo search algorithm leads to better quality clusters. Weighted temporal pattern mining is performed on these clusters to identify the effective temporal patterns which consider knowledge about the patterns having low frequency but high weight in a database which undergoes time based updates. Implementation of the proposed technique is carried out using MATLAB platform and its performance is evaluated using weather forecast dataset.

KEYWORDS-Time series database, Dimensionality reduction, Modified Adaptive Fuzzy C Means (MoAFCM), Cuckoo search optimization, Weighted temporal pattern mining.

I. INTRODUCTION

Data mining is the extraction of knowledge from huge databases and is one of the most important research areas in the recent years [1]. Mining performed on time dependent databases is called as temporal mining. This can be performed on time series databases where regularities in the data are identified based on time [2] [3] [16]. Each data item is associated with start time and end time which specifies the interval during which the concerned data item is valid [21] [22] [23] [24]. The threshold values of support and confidence are specified to mine temporal association rules of interest to the user [4] [5].

Clustering is one of the methods of data mining. Clustering is the concept by which data items which are similar are collected together in the same group [25]. That is, the objective of clustering is to group similar data together and high intergroup similarity. This is achieved on a temporal database initially, on the basis of time. In the recent years data is represented in a highly complex form. That is, each data item involves many attributes resulting in highly complex data. Temporal mining on complex data is tedious and difficult. Therefore it is necessary to reduce the data considered for mining. Fuzzy clustering algorithm is the most preferred clustering algorithm. Fuzzy clustering partitions the data into groups called clusters. Membership values are assigned to the data points with regard to each cluster. Fuzzy c means algorithm is a widely used classical clustering algorithm. Fuzzy c means (FCM) clustering is a well known method for clustering. It performs fuzzy partitioning in such a way so as to specify that a data point can belong to all groups or clusters with various membership degrees ranging between 0 and 1. But it has to be modified to be suitable for temporal mining on data with many dimensions [17] [18] [19] [20]. Therefore other optimistic ways of performing clustering on temporal data with dimensionality reduction must be considered. Dimensionality reduction can be considered as a process in which a single point represents a group of points. Only the relevant attributes are considered for further processing and the other attributes are considered non-relevant [25]. Optimization of the clustering process leads to good quality clusters. Clusters of high quality represent the least intergroup similarity and the highest intragroup similarity. Also searching and analyzing the data for similarity is an important task in clustering which can be optimized. Misclustering can be avoided by optimizing the clustering process. The factors that can be considered for optimization include time taken for clustering, memory usage and the computation complexity involved. High quality clusters lead to efficient frequent patterns and association rule mining [26] [27] [28].

Association rule mining is a method which identifies the frequent itemsets, which are preferred by the user [6]. The basic algorithm that was used for identifying the frequent itemsets was the Apriori algorithm. Since this algorithm required scanning of the original database large number of times, many algorithms came into existence to overcome the drawbacks of this algorithm. Frequent itemset mining on temporal dataset is difficult since a temporal database is incremental in nature [7] [8] [9]. That is, as time evolves new data are added to the database and the old frequent itemsets have to be updated according to the current state of the temporal database [10] [11]. Still more, if the database contains high dimensional data, the process of association rule mining becomes complex. In order to reduce the complexity, it is necessary to go for dimensionality reduction methods which result in data with less number of attributes which makes it suitable for association rule mining on a temporal database. Moreover it is necessary to consider the new records added to the temporal database every time mining is performed [12]. Only then the resulting frequent itemsets will be accurate and efficient. Therefore it is highly necessary for choosing an efficient mining algorithm which considers all the above mentioned factors for performing frequent itemset mining on a time series database [13] [14] [15].

Weighted temporal pattern mining discovers more important knowledge compared to conventional
frequent pattern mining. The drawback of frequent itemset mining is that, only the frequency of occurrence of data is considered for identifying the frequent patterns. By performing weighted pattern mining, knowledge about patterns which are frequent and also important is also considered for identifying the frequent patterns [29].

In our proposed weighted temporal pattern mining method, initially data transformation is performed on the time series database values to convert the numerical data into fuzzy data suitable for clustering. The MoAFCM algorithm which is a combination of FCM and Cuckoo search algorithm is applied on the transformed data. The FCM algorithm performs dimensionality reduction and clustering is performed by a combination of both FCM and Cuckoo search optimization algorithm resulting in optimized clusters of better quality. Then weighted temporal pattern mining is performed on these clusters, and effective temporal patterns are mined. The remaining sections of the proposed work are as follows: Section II reviews the research work related to the proposed technique. Section III explains the proposed technology, Section IV shows the experimental results and Section V concludes the paper.

II. LITERATURE SURVEY
A. PCFA: Projected Cluster Mining in High Dimensional data using modified FCM

Clustering of high dimensional data was a difficult process. Illango Murugappan et al. [25] have proposed a clustering method for high dimensional data by modifying the FCM clustering algorithm to overcome the difficulties of the K-means and FCM algorithms when applied to high dimensional data. The Projected Clustering based FCM Algorithm utilized the standard FCM clustering algorithm for sub-clustering high dimensional data into reference centroids. The matrix containing the reference values was used for clustering by the modified FCM algorithm. The four main steps of the PCFA technique was: 1) Gridding of the dataset by FCM algorithm, 2) Performing relevant attribute analysis on the gridded dataset, 3) Identification and removal of outliers from the dataset, 4) Clustering of high dimensional data by modified FCM.

Gridding reduced the time consumed in searching the dataset and reduced the space requirement. The dataset was initially divided into a set of blocks and FCM algorithm with a predefined number of clusters was applied to perform the gridding process on each block. The data points were replaced by reference values generated by the FCM algorithm. The centroid values of each cluster formed by the FCM algorithm represented the reference values. Hence the original data set was represented by a set of reference values. In relevant attribute analysis, projected clustering method was used to identify the cluster and the relevant attributes. A subset of the centroids was called as the projected cluster. The relevant and irrelevant attributes were identified. An m x n matrix was used for processing where centroids represented rows and attributes represented columns. Outlier detection removed non relevant data from the resulting matrix. Clustering of high dimensional data by modified FCM performed clustering of reference values by projected clustering. The number of clusters and reference values were given as input and the objective function of FCM was subject to optimization iteratively to obtain the required output. The modified FCM algorithm when evaluated in terms of clustering accuracy, memory usage and computational time was found to be better than the standard FCM algorithm and the Projected clustering technique that used K-means algorithm(PCKA).

B. Weighted fuzzy clustering algorithm for high dimensional data streams

Diksha Upadhay et al. [27] have proposed a dimension reduced weighted fuzzy clustering algorithm. This optimized fuzzy clustering algorithm could be used for data sets with higher dimension and also continuously arriving streaming behavior.

C. Clustering of high dimensional data using cuckoo search optimization algorithm

Priya Vajjayanthi et al. [28] have proposed a method for clustering of documents. There were a number of possible ways in which the documents can be clustered and this made the problem to be a combinatorial optimization problem. The cuckoo search optimization algorithm had been applied to solve the problem of document clustering.

D. Weighted frequent pattern mining

Ahmed C.F., et al. [29] have proposed a method for mining weighted frequent patterns from databases which were incremental in nature. Weighted pattern mining assigned different weights to the items in the transaction by which it was capable of discovering important knowledge when compared to the conventional methods of frequent pattern mining. By this method hidden knowledge about the patterns which have the least frequency but was of much importance were considered for identifying the frequent patterns. IWFP (Incremental Weighted Frequent Pattern Tree based on Weight Ascending order) tree structure and an algorithm IWFPWA for incremental and interactive weighted frequent pattern mining had been proposed. It avoided unnecessary computations on database updates or change of mining threshold by using the data structures and mining results obtained previously. Single scan of the database was sufficient for handling the data which were added to the database on temporal grounds without any repetitive work. The items were arranged in ascending order of the weight by which the item with the highest weight was placed at the bottom. This helped in effective generation of candidate patterns.

The proposed method which performs weighted temporal pattern mining in combination with modified AFM (MoAFCM) is compared with the existing PCFA(Projected clustering of high dimensional data using modified FCM algorithm) in terms of cluster accuracy, memory usage and computation time. It is found that the proposed method has better performance than the existing method which is discussed in detail in section IV of this paper.

III. PROPOSED WORK

In our proposed work, the numerical data of the original dataset which are high dimensional data are converted into fuzzy data. These fuzzy data form the result of data transformation. The transformed data is provided as input to the modified AFCM technique which is a combination of both FCM and cuckoo
search optimization algorithm. FCM performs dimensionality reduction over the input data and also clusters it along with cuckoo search optimization algorithm. Cuckoo search algorithm is used in combination with FCM to form and optimize the clusters formed by FCM algorithm. The MoAFCM algorithm produces optimized clusters which are of better quality. Frequent pattern mining is performed on these clusters using weighted temporal pattern mining which results in efficient and accurate frequent patterns. The process of data transformation, dimensionality reduction, clustering and weighted temporal pattern mining on a time series database is discussed as follows. Fig. 1 shows the block diagram for weighted temporal pattern mining technique using dimensionality reduction and optimized clustering.

A. Data Transformation
The database (D) is converted into fuzzy database (D') which has fuzzy data only. From the database D, data \( d \in D \) is taken with size of \( i \times j \) where \( i \in m \) and \( j \in n \) and the max (d), min(d) are computed. The median value is calculated as

\[
\text{Med}(d) = \frac{\text{max}(d) + \text{min}(d))}{4}
\]  

Four values are chosen in order to convert the numerical data into itemset data namely, Very Low, Low, High and Very High. Consider the numerical data as X.

- \( \text{min}(d) > X < = \text{Med}(d) \) - Very Low (VL)
- \( \text{VL} > X < = \text{Med}(d) \times 2 \) - Low (L)
- \( L > X < = \text{Med}(d) \times 3 \) - High (H)
- \( H > X < = \text{max}(d) \) - Very High (VH)

Based on the above conditions, the numerical data X present in the database (D) are converted into fuzzy data. This is provided as input to the proposed MoAFCM technique.

B. Modified Adaptive Fuzzy C Means Technique (MoAFCM)

i) Dimensionality reduction using the standard clustering method

Dimensionality reduction is performed by FCM. FCM is same as the K-means classifier which allows a single data item to belong to more than one cluster. A class membership value is specified to each data, depending on the resemblance of the data to a particular class in relation to other classes. The data are divided into blocks based on time and FCM is applied to each block with a predefined fixed number of clusters which is dependent on the number of time zones chosen. The relevant data for each cluster is selected by FCM algorithm. A centroid value is chosen for each cluster related with the corresponding time zone, and these centroid values are the reference points for the corresponding clusters. A projected cluster which contains only a subclass of the centroid points is considered for reducing the attributes. The subclass of centroids is chosen based on the time attribute. The attributes apart from time which are applicable to the particular cluster are the relevant attributes and the others are irrelevant attributes. The computation is based on a three dimensional matrix where time is a common attribute to both, the relevant attributes and the centroids. In the matrix the attributes represent columns and centroids represent rows. The irrelevant data points are removed using outlier detection.

ii) Cluster optimization using Cuckoo search algorithm

The resulting clusters after dimensionality reduction are optimized using cuckoo search algorithm in combination with FCM which is called as modified AFCM technique. Cuckoo search algorithm is a metaheuristic algorithm which was inspired by the breeding behavior of the cuckoos and is easy to implement. In cuckoo search, there are a number of nests. Each egg denotes a solution and an egg of cuckoo denotes a new solution. The new and better solution replaces the worst solution in the nest. A cuckoo is picked at random and it generates new solutions using levy flights. After that the generated cuckoo is evaluated using the objective function for determining the quality of the solutions.

The fitness function (\( f \)) of all the nests are computed by using the formula (2)

\[
O = \sum_{i=1}^{N} \sum_{j=1}^{m} m_{e_i}^x || x_i - C_{e_j} ||^2
\]  

Where \( m_{e_i}^x \) - membership of the \( i^{th} \) data to the \( j^{th} \) cluster, \( x_i \) - \( i^{th} \) data, \( C_{e_j} \) - centroid of the \( j^{th} \) cluster, \( N \) - Number of data points.

The quality of the solution is evaluated and a nest is selected arbitrarily. If the quality of new solution in the selected nest is better than the old solutions, it will be replaced by the new solution. Otherwise, the previous solution is kept as the best solution.

The worst nests are discarded based on their probability. The best solutions are ranked according to their fitness function. Then the best solutions are identified and marked as optimal solutions. This process is repeated until the maximum iteration is reached. Finally, the best clustered dataset is obtained. After the clustering process, the transaction id is given to each itemset (cluster) present in the obtained dataset.

C. Weighted Temporal Pattern mining

The clusters obtained from the modified AFCM technique are considered as transactions and are assigned transaction id. Weighted frequent pattern mining [29] is performed using a tree based weighted frequent pattern mining approach on the basis of time. Every item in the transaction is assigned a weight which denotes its significance. Weight of an item is a non-negative real number which is assigned to reflect the importance of each item in the transaction database.

For a set of items \( I = \{i_1, i_2, \ldots, i_l\} \), weight of a pattern \( P[i_1, i_2 \ldots i_l] \) is given as follows:

\[
\text{Weight}(P) = \sum_{i=1}^{l} \text{Weight}(x_i) / \text{length}(P)
\]
The weighted support of a pattern is defined as the resultant value of multiplying the pattern’s support with the weight of the pattern and is given as follows:

$$W_{\text{support}}(P) = \text{Weight}(P) \times \text{Support}(P) \quad (4)$$

A pattern is called a weighted frequent pattern if the weighted support of the pattern is greater than or equal to the minimum support threshold.

When new data is added to the time series database dynamically, weighted temporal pattern mining process is carried out only to those newly added data. This is performed using a tree based weighted frequent pattern mining approach based on the Incremental Weighted Frequent Pattern Tree with Weight Ascending order tree structure and the incremental and interactive weighted frequent pattern mining algorithm [29]. It avoids unnecessary computations on database updates or change of mining support threshold by using the data structures and mining results obtained previously. Single scan of the database is sufficient for handling the data which are added to the database on temporal grounds without any repetitive work. The items are arranged in ascending order of the weight by which the item with the highest weight is placed at the bottom. This helps in effective generation of candidate temporal patterns.

IV. EXPERIMENTAL RESULTS

Our proposed weighted temporal pattern mining technique has been validated by experiments with a weather forecast dataset. The dataset consists of 5 years of weather forecast data from the year 2011 to 2015. It is the district-wise weather forecast data with many attributes, recorded at various districts of Tamilnadu, India on daily basis. The data has been obtained from the Indian Meteorological Department, India. The proposed system has been implemented in MATLAB (Version R 2012a). Here we employ the weighted temporal pattern mining algorithm for the purpose of mining weighted frequent patterns from temporal databases. Initially the temporal database is transformed into fuzzy data. These fuzzy data are provided as input to the modified AFCM algorithm which is a combination of FCM and cuckoo search optimization technique. This performs dimensionality reduction and clustering, the result of which is optimized clusters. These clusters are considered as transactions based on time and are provided as input to the weighted temporal pattern mining algorithm.

The output of which are effective temporal patterns which is based not only on the frequency but also on the importance of the item denoted by the weight assigned to it. Also, whenever new transactions are added to the database, the process of frequent pattern mining is not repeated for the same set of data, instead only a single scan of the database is performed and previously stored information are used for mining the updated database. The main function of the algorithm is to minimize the computation time, memory usage. It also increases the cluster accuracy as well as the efficiency of the weighted temporal patterns obtained. The performance of the proposed technique is compared with that of existing FCM and PCFA methods. It is found that the proposed system has better performance in terms of computation time, memory usage, cluster accuracy and efficiency of the resulting temporal patterns.

A. Performance Analysis

The contents of the weather forecast dataset is in numerical form which cannot be processed for further operations. So it is converted into fuzzy data. The numerical dataset is converted into fuzzy data on weekly basis. Table 1 and Table 2 shown below display the sample weather forecast dataset containing numerical data and the corresponding fuzzy data respectively.

### TABLE 1 SAMPLE WEATHER FORECAST DATASET CONTAINING NUMERICAL DATA

<table>
<thead>
<tr>
<th>DISTRICT: CHENNAI</th>
<th>5Sep15</th>
<th>6Sep15</th>
<th>7Sep15</th>
<th>8Sep15</th>
<th>9Sep15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rainfall (mm)</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Max Temperature (°C)</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>Min Temperature (°C)</td>
<td>26</td>
<td>26</td>
<td>26</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>Total cloud cover (%)</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Min Relative Humidity (%)</td>
<td>86</td>
<td>86</td>
<td>86</td>
<td>86</td>
<td>86</td>
</tr>
<tr>
<td>Wind speed (knots)</td>
<td>65</td>
<td>65</td>
<td>65</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td>Wind direction (deg)</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

### TABLE 2 FUZZY DATA

<table>
<thead>
<tr>
<th>DISTRICT: CHENNAI</th>
<th>5Sep15</th>
<th>6Sep15</th>
<th>7Sep15</th>
<th>8Sep15</th>
<th>9Sep15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rainfall (mm)</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Max Temperature (°C)</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>Min Temperature (°C)</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Total cloud cover (%)</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Min Relative Humidity (%)</td>
<td>92</td>
<td>93</td>
<td>93</td>
<td>94</td>
<td>94</td>
</tr>
<tr>
<td>Wind speed (knots)</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Wind direction (deg)</td>
<td>180</td>
<td>180</td>
<td>180</td>
<td>180</td>
<td>180</td>
</tr>
</tbody>
</table>

From Fig. 2 below it is noted that the number of temporal frequent patterns obtained for various number of transactions using the MoAFCM algorithm is more than that of PCFA and FCM techniques. Thus the proposed technique extracted temporal patterns more efficiently from the time series database.
Fig. 3 shows that the number of frequent patterns obtained from a time series database for different values of support is more in the case of the proposed MoAFCM technique when compared with that of PCFA and FCM techniques. Since support is a user specified value, the MoAFCM technique is found to be better for a large range of support values.

From Fig. 4 below it can be found that the computation time taken by the proposed MoAFCM technique is less when compared with PCFA and FCM techniques for varied number of transactions.

From Fig. 5 it is found that the computation time taken by the proposed MoAFCM technique is less when compared to that of the existing PCFA and FCM techniques for different values of minimum support percentage.

As displayed in Fig. 6 the proposed technique performs better in terms of memory usage than the existing technique for weighted frequent pattern mining. It is found that the memory utilized by the proposed technique is less when compared with that of the existing technique. This is because of dimensionality reduction and usage of previous data structures by the proposed technique.

Fig. 7 displays that the accuracy of the clusters formed by the proposed MoAFCM technique are better than that of the existing PCFA method for various number of clusters. This is because optimized clusters are formed by MoAFCM which is the combination of FCM and the cuckoo search optimization algorithm.

Fig. 8 shows that the computation time taken for weighted temporal pattern mining when using MoAFCM is less when compared to that of the existing PCFA method for varying number of transactions. This is because of dimensionality reduction prior to clustering and single scan of the transaction database during the mining process.

From Fig. 9 below it is found that in the case of the proposed weighted temporal pattern mining approach in combination with MoAFCM technique, the number of weighted temporal patterns is more for various weighted support values, when compared to that of the existing weighted temporal pattern mining algorithm.
From Fig. 10 it is found that the weighted temporal rules resulting from the proposed weighted temporal pattern mining approach in combination with MoAFCM are more accurate than that obtained by the existing method of weighted temporal pattern mining for varying number of transactions. This is because of considering the importance of the item (that is, weight) along with its frequency for mining the frequent patterns.

**Fig.9. Comparison of the performance of the proposed and existing technique for weighted temporal pattern mining in terms of no. of weighted temporal patterns by varying the weighted support value**

- Fig.10. Comparison of the performance of the proposed and existing technique for weighted temporal pattern mining in terms of accuracy of weighted temporal rules by varying the number of transactions

It is found from Fig. 11 that the accuracy of the weighted temporal rules obtained by the proposed technique for weighted temporal pattern mining is more than that obtained by the existing technique for weighted temporal pattern mining for various weighted support values. This is because the mining process takes into account the importance of the item (weight assigned) along with its frequency for identifying the frequent patterns.

**Fig.11. Comparison of the performance of the proposed and existing technique for weighted temporal pattern mining in terms of accuracy of weighted temporal rules by varying the weighted support value**

**V. CONCLUSION**

In this paper weighted temporal pattern mining on a time series database using MoAFCM algorithm is proposed. The MoAFCM technique is a combination of FCM and cuckoo search optimization algorithm. Dimensionality reduction is performed by FCM and optimized clustering is performed by a combination of FCM and cuckoo search optimization algorithm. Due to dimensionality reduction prior to clustering, MoAFCM technique leads to reduced computation and time complexity. Also from the point of view of cluster efficiency, the clusters resulting from MoAFCM technique are more efficient when compared to that of the existing PCFA and FCM techniques. Since time is a major component in temporal databases, the MoAFCM algorithm is better suited for time based applications and weighted temporal pattern mining with less memory utilization can be performed on temporal databases which undergo updates in a time based manner. Thus weighted temporal pattern mining in combination with MoAFCM with dimensionality reduction is well suited for time based datasets and can be efficiently utilized in real time environments.

**REFERENCES**


