An Efficient Data Clustering Technique Using Flexible and Accurate Motif Detector Algorithm

S. Vasanthakumar
Research Scholar, Department of Computer Science, K.S.G College of Arts and Science, Coimbatore-15, Tamil Nadu, India.

N. Siva Kumar
Assistant Professor & Head, Department of Information Technology, K.S.G College of Arts and Science, Coimbatore-15, Tamil Nadu, India.

Abstract – The this paper we present a new Flexible and Accurate Motif Detector algorithm (FLAME), this algorithm is a most flexible one to build the tree generation of probability suffix order, find similarity and merging the sequence data. The usage of FLAME, it is currently possible to find the mine datasets that would have been most highly used to find the sequence pattern which is difficult with already existing tools. To reduce the size of the datasets is to use the sequences of character store in a particular memory with the same character in the tree based approaches to reduce the memory size in the existing tools. The experiment result shows that FLAME algorithm has good performance when compared with other methods.

Index Terms – Data mining, Clustering, FLAME, Accuracy, K-Means.

1. INTRODUCTION

Data mining is a system of procedure for search the large amount of data in the pattern for different techniques and that is directly related to various kinds of new concepts is directly related to computer science. Despite this, it can be used with a number of older computer techniques such as pattern recognition and statistics. The goal of the data mining techniques that handles large number of applications. It is commonly using for organisations or businesses to process the needs to complete the trends or patterns. The main purposes of the data mining are to refer the information of data was not previously known to retrieve. Generally, data mining (sometimes called data or knowledge discovery) is the process of analysing data from different perspectives.

Data mining software is one of the analytical tools for analysing data in the number of data. The data mining is the techniques to find the data from many different angles or dimensions and to categorise it and analysis the relationships. It is used for technical purposes to finding the patterns or correlations among large number of fields in relational database. Data mining is the techniques to be increased for both the public and private sectors. The data mining concepts can be used for industrial areas such as insurance, banking, retailing and medicine is commonly used for data sectors to calculate to reduce cost, increase sales and development of resources. In public sector data mining applications basically used for detect fraud and waste but also mainly used for the purposes of measuring the data size and improve the performances.

2. METHODOLOGY

The proposed system present a new algorithm called Flexible and Accurate Motif Detector (FLAME).FLAME is a flexible suffix tree based algorithm that can be used to find frequent patterns with a variety of definitions of motif (pattern) models. It is also accurate, as it always finds the pattern if it exists. In addition, a clustering algorithm is proposed to find the group relationships for query and data aggregation efficiency.

First, since the clustering algorithm itself is a centralized algorithm, the dissertation further considers systematically combining multiple local clustering results into a consensus to improve the clustering quality and for use in the update based tracking network.

Second, when a delay is tolerant in the tracking application, a new data management approach is required to offer transmission efficiency, which also motivates this study. The dissertation defines the problem of compressing the location data of a group of moving objects as the group data compression problem.

3. K-MEANS CLUSTERING

Given a data set of data samples, a desired number of clusters, k-mean and a set of k initial starting [19] points, the k-means clustering algorithm finds the desired number of distinct clusters and their centroids. A centroid is described as the point whose coordinates are obtained by computing the average of each of the coordinates (i.e., feature values) of the points of the jobs allocated to the cluster. Formally, the k-means clustering algorithm pursues the following steps.

Step 1: Choose a number of desired clusters, k.

Step 2: Choose k starting points to be used as preliminary estimates of the cluster centroids. These are the initial starting values.

Step 3: To examine the assign it to the cluster whose centroid is nearest to it and each point in the data set.

Step 4: When each point is assigned to a cluster, recomputed the new k centroids.

Step 5: To repeat steps 3 and 4 till no point changes its k-means cluster assignment or until a maximum number to be passes through the data set for performed.

The z-score equation is

$$F^*{}_{im} = \frac{F_{im} - \mu_m}{\sigma_m}$$

4. FLAME ALGORITHM

The FLAME algorithm, which can be used to find (L, M, s, k) motifs. For ease of exposition, we explain the algorithm using an (L, d, k) model, and then describe how we extend it to the full-fledged (L, M, s, k) model. To recall the (L, k, d) gets the original results of the motif sequence of the string length L that occurs n number times in k in the dataset, the finding the each occurrence with the Hamming distance of d to define the model string. Given, A C B BBCACCB CCB BCACCB CACCB null ACCB CB B 2 3 3 8 String = ABBCACCB. A count suffix tree on the string ABBCACCB.

The counts are indicated inside the node. However, this approach might miss motifs as the model string might not actually occur in the dataset even once. To suppose for illustration of the string ABCDEF is the true motif. Assume that we are looking for a (6, 2, 3) pattern, and that the instances of this pattern in the dataset are FFCDEF, ABFFEF, and ABCDAA.

FLAME Algorithm Code

```
FLAME (modelTree, dataTree, l, d, k)
1. Model = modelTree.FirstNode ()
2. While (model 6= modelTree.LastModel ())
3. Evaluate Support(model,dataTree)
4. If (isValid (model)) Print “Found Model:”, model
5. Else If(model.support () < k)
6. ModelTree.PruneAt (model)
7. Model = NextNode (model,modelTree)
8. End While
9.End
```

Each instance is at a distance of 2 from the model ABCDEF, but the distance between any two instances is 4. If we consider only instances from the dataset (which need not contain ABCDEF), then we will not find the motif. The approach willtakethe FLAME and to explores the space of all possible methods of models. In order to carry out this exploration in an efficient way, we first construct two suffix trees: a count suffix tree on the actual dataset (called the data suffix tree), and a suffix tree on the set of all possible model strings (called the model suffix tree).

5. SEQUENCIAL PATTERN MATCHING

Let α be a sequential pattern matching in a data mining for sequence of S, the α-projected database, denoted as S|α, is the collection of sequential pattern suffixes of S with regard to prefix α. Algorithm 1 outlines the mining process. Assuming that the current pattern is frequent, the algorithm extends it by appending one base at a time (Line 3), and constructs the corresponding projected database (Line 4). If the extended pattern is frequent and closed (Algorithm 2), then the algorithm recursively calls itself with the extended pattern (Line 8). Therefore, the current pattern is always closed. If the current pattern cannot extend to any frequent pattern (Line 9), it is maximal according to Definition 1. And if the pattern is long enough, it will be saved with its projected database (Line 10).To check whether a frequent pattern is closed, we adopt the method proposed, which is outlined in Algorithm 2.

6. DATASET DESCRIPTION

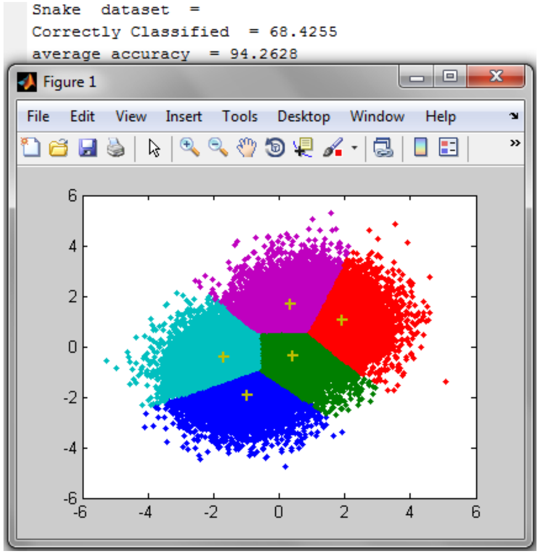
The characteristics of the evaluated data sets are summarized in the following. Synthetic data sets. We also exploited a synthetic data set generator to evaluate algorithm performance and scalability. The data generator is based on the snake dataset generator. It allows generating transactional synthetic data sets by setting (i) the dataset cluster, (ii) the correctly classified, and (iii) the average accuracy. To assign weights to data generated by the snake dataset generator we integrated a synthetic weight generator. The newly proposed data generator version may assign to each data item a weight according to two different distributions, chosen as representative among all the possible data distributions, i.e, the uniform data distribution and the Possible to distribution. When not otherwise specified, in the following experiments item weights are selected in the range [1, 100].

7. COMPRESSION ALGORITHM

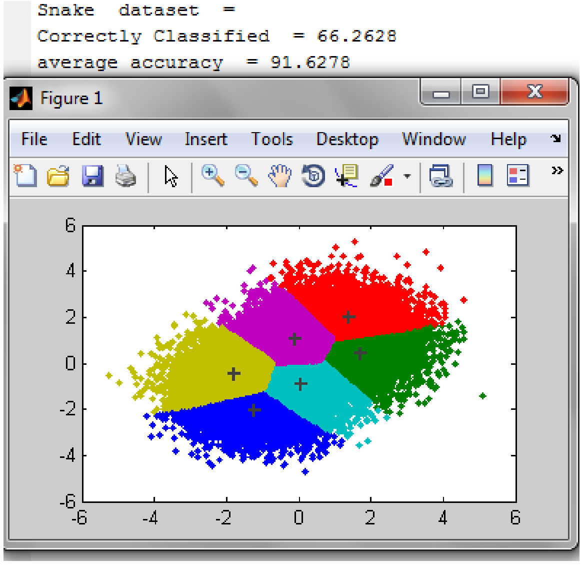
The algorithm trims and prunes more items when the group size is larger and the group relationships are more distinct. Besides, in the case that only the location center of a group of objects is of interest, the approach can find the aggregated value in the phase, instead of transmitting all location sequences back to the sink for post-processing.To compress the location sequences for a group of moving objects, the proposed system processes the Merge algorithm.

8. RESULTS

The cluster algorithm can be the comparison of snake datasets and leukemia datasets for the clustering of the data and comparisons and correctly classified of and finding the average accuracy rate can be increased by the cluster from 1 to 5 and 1 to 11 of various rates of the results.



```
Snake  dataset  =
Correctly Classified  = 68.4255
average accuracy  = 94.2628
```



```
Snake  dataset  =
Correctly Classified  = 66.2628
average accuracy  = 91.6278
```

Table:1 Motif Pattern in Snake Datasets

TIME IN SEC	LOG SCALE	MOTIF
0.6	6	10
0.8	8	10
1	10	10
50	10	20
50	12	20
800	12	30

On comparing of various methods to calculate the datasets for finding the different pattern results of YMF, Random projection algorithm which make possible for getting the results of comparing the K-means clustering method. It gives the higher result of snake datasets and leukemia datasets the leukemia datasets which in turn give the higher results of the patterns of the K-means clustering in the Flame projection algorithm and K-Means algorithms.

9. CONCLUSION

The experimental results show that the proposed compression algorithm leverages the group movement patterns to reduce the amount of delivered data effectively and efficiently. The dissertation eliminates the difficulties in the existing system. It is developed in a user-friendly manner. The system is very fast and any transaction can be viewed or retaken at any level. Error messages are given at each level of input of individual stages.

REFERENCES

[1] Ravi Ranjan and G. Sahoo, "A New Clustering Approach For Anomaly Intrusion Detection", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.2,pp.29-38, 2014.

[2] Srinivas Sivarathri1 and A.Govardhan, Experiments On Hypothesis "Fuzzy K-Means Is Better Than K-Means For Clustering", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.5, September 2014.

[3] K.Prabha and K.Rajeswari, "A Hybrid Approach for Data Clustering Using Data Mining Techniques", IJCSMC, Vol. 3, Issue. 11, November 2014, pp.81– 88

[4] Mark J. Van der Laan and Katherine S. Pollard "A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap"

[5] Aditi purohit1, Hitesh Gupta, "Hybrid Intrusion Detection System Model using Clustering, Classification and Decision Table",IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727Volume 9, Issue 4 (Mar. - Apr. 2013), pp.103-107.

[6] Gurjit Singh and Navjot Kaur, "Implementation of Hybrid Clustering Algorithm with Enhanced K-Means and Hierarchal Clustering" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2013.

[7] Vivek K. Kshirsagar, Sonali M. Tidke& Swati Vishnu, "Intrusion Detection System using Genetic Algorithm and Data Mining".

[8] Manish Somani, RoshniDubey," Hybrid Intrusion Detection Model Based on Clustering and Association", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 3, Issue 3, March 2014.

[9] ChapkePrajkta P., Raut A. B. "Hybrid Model For Intrusion Detection System", International Journal of Engineering and Computer Science ISSN:2319-7242. Volume1 Issue 3 Dec 2012 pp. 151-155.

[10] Chandrashekhar Azad and Vijay Kumar Jha, "Data Mining based Hybrid Intrusion Detection System", Indian Journal of Science and Technology, Vol 7(6), 781–789, June 2014.