Simultaneous Pattern and Data Clustering Using Modified K-Means Algorithm

M.Pramod Kumar
Vignan University,
Vadlamudi, Guntur,
Andhrapradesh.

Prof K V Krishna Kishore
Head of the Dept, CSE,
Vadlamudi, Guntur,
Andhrapradesh.

Abstract-- In data mining and knowledge discovery, for finding the significant correlation among events Pattern discovery (PD) is used. PD typically produces an overwhelming number of patterns. Since there are too many patterns, it is difficult to use them to further explore or analyze the data. To address the problems in Pattern Discovery, a new method that simultaneously clusters the discovered patterns and their associated data. It is referred to as "Simultaneous pattern and data clustering using Modified K-means Algorithm". One important property of the proposed method is that each pattern cluster is explicitly associated with a corresponding data cluster. Modified K-means algorithm is used to cluster patterns and their associated data. After clusters are found, each of them can be further explored and analyzed individually. The proposed method reduces the number of iterations to cluster the given data. The experimental results using the proposed algorithm with a group of randomly constructed data sets are very promising.

Index Terms- Pattern Discovery, Contingency table, and Chi-Square test.

1 INTRODUCTION

The process of grouping a set of physical objects into classes of similar objects is called **Clustering**. A **Cluster** is a collection of data objects that are similar to one another with in the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group in many applications. Dissimilarities are accessed based on the attribute values describing the objects. Often

distance measures are used. Clustering has its roots in many areas, including data mining, statistics, biology and machine learning. Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabelled data. A loose definition of clustering could be "the process of organizing objects onto groups whose members are similar in some way". Cluster analysis is an important human activity. Early in childhood, one leans how to distinguish between cats and dogs, or between animals and plants, by continuously improving subconscious clustering schemes. Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing, and market research. By clustering, one can identify dense and sparse regions and therefore, discover overall distribution patterns and interesting correlations among data attributes. The patterns belonging to the same cluster having the same label. In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer [8].

The basic idea of PD [4] can be illustrated by a simple XOR problem with three binary variables: A, B and C= A XOR B. Suppose that we want to check whether or not the occurrences O of the compound event [A=T, B=T, C=F] is just a random happening. Given the observed frequency of occurrences O of the compound event, if we could estimate its expected frequency of occurrences e under the random assumption. A compound event is called an event association pattern or simply a pattern, if the difference (O-e) is significant enough to indicate that the compound event is not a random happening. PD is a useful tool for categorical data analysis. The patterns produced are easy to understand. Hence it is widely used in business and commercial applications. PD typically produces an overwhelming number of patterns. The scope of each pattern is very difficult and time consuming to comprehend. There is no systematic and objective way of combining fragments of information from individual patterns to produce a more generalized form of information. Since there are too many patterns, it is difficult to use them to further explore or analyze the data.

To address the problems in Pattern Discovery, We propose a new method that simultaneously clusters the discovered patterns and their associated data. It is referred to as "**Simultaneous pattern and data clustering using Modified k-Means algorithm**". One important property of the proposed method is that each pattern cluster is explicitly associated with a corresponding data cluster. To effectively cluster patterns and their associated data, several distance measures are used. Once a distance measure is defined, existing clustering methods can be used to cluster patterns and their associated data. After clusters are found, each of them can be further explored and analyzed individually. The above procedures for handling a large number of patterns

are based on a divide-and-conquer approach. In the divide phase, patterns and data are simultaneously clustered and in the conquer phase, individual clusters are further analyzed

2 LITERATURE STUDY

Agrawal and Srikanth [1] developed association rule mining for transaction databases. It is the process of finding frequent patterns with in the data of some database. Mining rules are useful to gain information, knowledge, etc. An association rule is of the form A=>B where A, B included in I and (A ^B=ø).Performance is measured via support and confidence. (Where I is an item set). Support:- The support of a rule, A=>B, is the percentage of transactions in DB, the DB containing both A and B. Support is an actual frequency Confidence, c(X->Y) = σ (XUY)/ σ(X).They have used apriority property. Apriori property: - All non-empty subset of a frequent item set must also be frequent. An item set is said to be frequent if it satisfies the minimum support threshold.

	Tea=Y	Tea=N	Row Sum
Coffee=Y	20	70	90
Coffee=N	5	5	10
Col. Sum	25	75	100

Table-1 Contingency Table of the Purchase of Tea and Coffee

Brin [3] proposed the use of chi-square statistics to detect Correlation rules from contingency tables. One way of measuring the correlation is Corr a, b =p (AUB) /p (A) p (B). If the result is equal to 1, then both A and B are independent. If the resulting value is greater than 1, then A and B are positively correlated, else A and B are negatively correlated. The fact is that we calculate the correlation value indeed, but we could not tell whether the value is statically significant. So, Bring introduced the chi squared for independence. Brin takes into account all possible combinations of the presence and absence of various attributes. The **chi-squred statistic** is defind as**: X2 = *(O - E) 2 /E*** Where O is observed frequency and E is expected frequency .If the X2 is equal to 0, then all the variables are really independent. If it is larger than a cutoff value at one significance level, then we say all the variables are dependent (correlated), else we say all the variables are independent. When the contingency table data is sparse, correlation rule is less accurate. The chi-square test does not give us much information about the Strength of the relationship or its *substantive significance* in the Population. The chi-square test is also sensitive to small expected frequencies in one or more of the cells in the table. To address the problem in pattern discovery, Wong and Li [2] have proposed a method that simultaneously clusters the discovered patterns and their associated data. In which pattern induced data clusters is introduced. It relates patterns to the set of compound events containing them and makes the relation between patterns and their associated data explicit. Pattern induced data clusters defined are constants. That is each attribute has only one value in the cluster. Since each pattern can induce a constant cluster, the number of constant clusters is overwhelming. To reduce the number, it is desirable to merge clusters. Let us say two clusters I (i), I (j) are two clusters. The merged data cluster of I (i) and I (j) is the union of their matched samples and matched attributes. When two data cluster are merged, the corresponding patterns including them are simultaneously clustered. Distance measure is used. Once a measure is defined, existing clustering methods can be used. They have used hierarchical agglomerative approach.

3 THE CLUSTERING ALGORITHM*:*

Once distance measure is defined, many clustering algorithms can be applied to clustering algorithms can be applied to clusters patterns. In this work, Modified k-Means algorithm [17] has been used.

3.1. Modified K-Means Clustering Algorithm:

Let $D = \{d(j) | j = 1, n\}$ be a data set having K clusters, $C = \{ci | i = 1, K\}$ be a set of K centers
And $Sj = \{d(j) | d(j)$ is member of cluster k$\}$ be the set of samples that belong to the jth cluster. Conventional K Means algorithm minimizes the following function which is defined as an objective function

$$Cost(D,C) = \sum_{j=1}^{n} dist(d^{(j)}, c_k) \qquad \textbf{(1)}$$

Where dist ($d^{(j)}$, ck) measures the Euclidean distance between a points $d^{(J)}$ and its cluster center c_K. The k-means algorithm calculates cluster centers iteratively as follows:
1. Initialize the centers in c_K using random sampling;
2. Decide membership of the points in one of the K clusters according to the minimum distance from cluster center criteria;
3. Calculate new c_K centers as:

$$c_k = \frac{\sum_{d^{(j)} \in S_k} d^{(j)}}{|S_k|} \qquad (2)$$

Where $|S_k|$ is the number of data items in the kth cluster;
4. Repeat steps 2 and 3 till there is no change in cluster centers.

Instead of using centers found by (2) every time, our proposed algorithm calculates the cluster centers that are quite close to the desired cluster centers. The proposed algorithm, first divides the data set D into K subsets according to some rule associated with data space patterns, then chooses cluster centers for each subset.

3.2. Outline of the proposed algorithm

Consider a data set D = {d (j) = (d (j)
1 . . . . d (j) in Rm and K is predefined number of clusters. Bellow is the outline of a precise cluster centers initialization method.

Step1.Dividing D into K parts

$D = \bigcup_{k=1}^{K} S_{k,} S_{k1} \bigcap S_{k2} = \varnothing, k_1 \neq k_2$ according to data patterns;

Step2.clculate new C_k centers as the optimal solution of

$$\min z = \sum_{d^{(j)} \in S_k} \left\| x - d^{(j)} \right\|, x = (x_1, ..., x_m) \in R^m . \quad \textbf{(3)}$$

Where $\left\| \bullet \right\|$ denotes the 2-norm.

Step3.Decide membership of the patterns in each one of the K- clusters according to the minimum distance from cluster center criteria.

Step4. Repeat steps 2 and 3 till there is no change in cluster centers.

4. EXPERIMENTAL RESULTS

Dataset is a set of data items. Data items are stored in the database; they can be represented in the form of data points in a two-dimensional space. In modified K-Means algorithm, user can enter the number of data points In modified K-Means algorithm user can specify the number of data points in advance; Where K means the number of clusters, as we want.

Dataset

A Data Set is a set of items. It is usually represented in tabular form. It is roughly equivalent to a two dimensional spread sheet or data base table. The rows of a table represent the members of a data set. The columns of a table represent the features or attributes of the data items. A simple database [18] containing 17 Boolean-valued attributes. The "type" attribute appears to be the class attribute. Here is a breakdown of which animals are in which type:

Zoo Dataset

1. Class# Set of animals:

1 (41) Aardvark, Antelope, Bear, Boar, Buffalo, Calf, Cavy, Cheetah, Deer, Dolphin, Elephant, Fruitbat, Giraffe, Girl, Goat, Gorilla, Hamster, Hare, Leopard, Lion, Lynx, Mink, Mole, Mongoose, Opossum, Oryx, Platypus, Polecat, Pony, Porpoise, Puma, Pussycat, Raccoon, Reindeer, Seal, Sealion, Squirrel, Vampire, Vole, Wallaby, Wolf

2 (20) Chicken, Crow, Dove, Duck, Flamingo, Gull, Hawk, Kiwi, Lark, Ostrich, Parakeet, Penguin, Pheasant, Rhea, Skimmer, Skua, Sparrow, Swan, Vulture, Wren

3 (5) Pitviper, Seasnake, Slowworm, Tortoise, Tuatara

4 (13) Bass, Carp, Catfish, Chub, Dogfish, Haddock, Herring, Pike, Piranha, Seahorse, Sole, Stingray, Tuna

5 (4) Frog, Frog, Newt, Toad

6 (8) Flea, Gnat, Honeybee, Housefly, Ladybird, Moth, Termite, Wasp

7 (10) Clam, Crab, Crayfish, Lobster, Octopus, Scorpion, Seawasp, Slug, Starfish, Worm

2. Number of Instances: 101

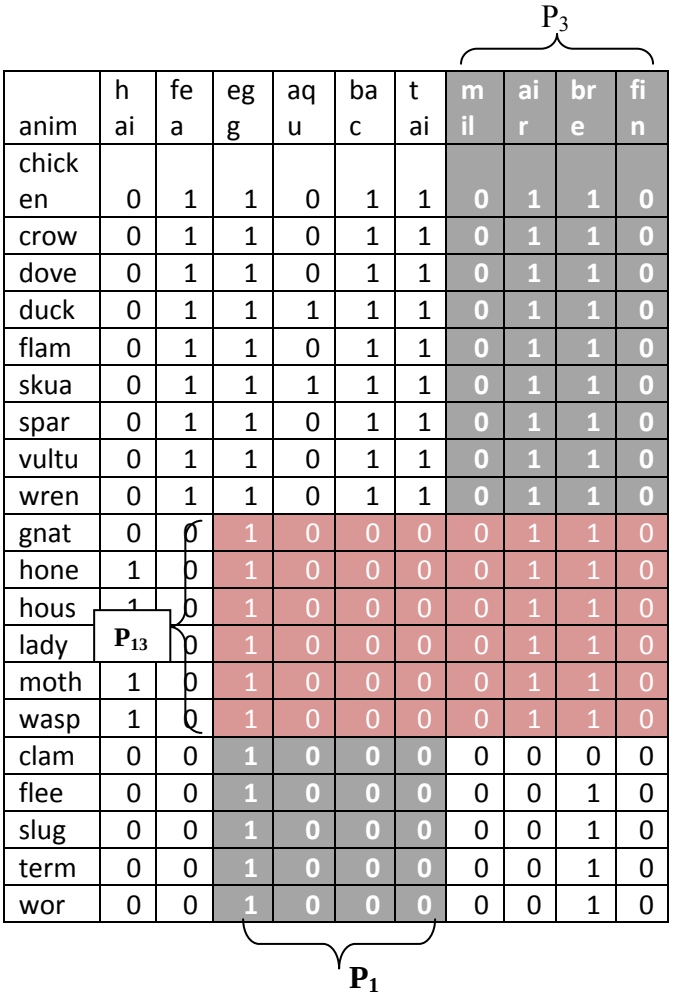
3. Number of Attributes: 18 (Animal Name, 15 Boolean Attributes, 2 Numeric)

4. Attribute Information: (Name of Attribute and Type of Value Domain)

1.	Animal attribute name	Unique for each instance
2.	Hair	Boolean
3.	feathers	Boolean
4.	eggs	Boolean
5.	milk	Boolean
6.	Airborne	Boolean
7.	Aquatic	Boolean
8.	Predator	Boolean
9.	toothed	Boolean
10.	Backbone	Boolean
11.	breathes	Boolean
12.	Venomous	Boolean
13.	fins	Boolean
14.	Legs	Numeric (set of values: {0, 2, 4, 5, 6, and 8})
15.	tail	Boolean
16.	Domestic	Boolean
17.	catsize	Boolean
18.	Type	Numeric (integer values in range [1,7])

Table.2 Data names and attributes

Once modified K-Means Algorithm is applied to the data points in the Two-dimensional space, the data points are divided into K-clusters based on the mean distance from the data point and the cluster centroids. Finally we get the K required number of clusters. The stopping criterion of pattern clustering depends on the measure that it uses. If dR and dRC are used, stopping criteria dR > 1 and dRC > 1 are available.



animal	hair	fea	egg	aqu	bac	tai	mil	air	bre	fin
chicken	0	1	1	0	1	1	0	1	1	0
crow	0	1	1	0	1	1	0	1	1	0
dove	0	1	1	0	1	1	0	1	1	0
duck	0	1	1	1	1	1	0	1	1	0
flam	0	1	1	0	1	1	0	1	1	0
skua	0	1	1	1	1	1	0	1	1	0
spar	0	1	1	0	1	1	0	1	1	0
vultu	0	1	1	0	1	1	0	1	1	0
wren	0	1	1	0	1	1	0	1	1	0
gnat	0	0	1	0	0	0	0	1	1	0
hone	1	0	1	0	0	0	0	1	1	0
hous	1	0	1	0	0	0	0	1	1	0
lady	1	0	1	0	0	0	0	1	1	0
moth	1	0	1	0	0	0	0	1	1	0
wasp	1	0	1	0	0	0	0	1	1	0
clam	0	0	1	0	0	0	0	0	0	0
flee	0	0	1	0	0	0	0	0	1	0
slug	0	0	1	0	0	0	0	0	1	0
term	0	0	1	0	0	0	0	0	1	0
wor	0	0	1	0	0	0	0	0	1	0

Fig: 1 zoo data set [15]

4.1. Distance Measure

Let r_i be the number of samples matched by x_j^{si} and r_j is the number of samples matched by x_j^{sj} that is $r_i=|m(i)\backslash m(J)|$ & $r_j=|m(j)\backslash m(i)|$. Let r_{ij} be the number of samples matched by both x_j^{si} and x_j^{sj} and .That is $r_{ij}=|m(i) \wedge m(j)|$. The distance is defined as $d_T(i, j) = r_i+r_j$. Where d_T is the Toivonen distance.

Example: From the above data set $r_i=11$, $rj=16$, $rij=6$.$d_T(i, j) =11+16=27$.Toivonen distance d_T tends to give higher values for rules that are matched by more sample.

To address this problem, normalized distance $D_g(i, j) =1- r_{ij}/ (r_i+r_j+r_{ij})$.From the given data set $r_{ij} =6$, $r_i=11$, $rj=27$.$d_g(i, j) =1- 6/ (11+27+6) =0.8181818$.

We can find the ratio of matched samples $d_r(i, j)=r_i+r_j/r_{ij}=27/6=4.5$.If d=1, then the number of different samples is same as the number of common samples. It can be used as a natural threshold for stopping a clustering algorithm.

If dr>1, then there is s more dissimilarity between the two patterns. The above measure does not give special consideration to the attributes where the patterns share or differ as an illustration consider the two pairs of patterns and x_i^{si},x_j^{sj} and x_p^{sp},x_q^{sq}.

Let c_{ij} is the number of attributes matched by both x_i^{si}, xjsj i.e. $|s_1 \wedge s_2|$.It seems more reasonable to consider that are similar, since they share certain attributes (cij>0).While and are not (cpq=0) one possible measure for considering both the matched samples and the matched attributes Where wc wr are the weights of the samples and the attributes respectively.

Note: If we consider the number of matched samples and matched attributes equally important, we may set to 0.5.Let us calculate

Example:

D (i, j) = (0.5) (11+27)/6+ (0.5) (4+4)/8
=0.5*6.3+0.5*1
=3.6

One problem of measure is that it does not consider the variation within the data cluster .To obtain good data clusters, we would like to minimize variations in the clusters.

4.2. SIMULTANEOUS PATTERNS AND DATA CLUSTERING

Suppose that there are set of patterns $\{x_1^{s1}, x_2^{s2}, x_3^{s3}$, and x4s4 $\ldots x_n^{sn}\}$. Then, the set of samples matched by a patterns x_i^{si} is devoted by $m(i) = \{x \varepsilon D/x \geq x_i^{si}\}$.

A pattern induced data cluster of a pattern x_i^{si} is a set of Compound events containing x_i^{si} and is represented by $I(i) = \{x^s \leq x/x \varepsilon m(i), s = s_i\}$

As an Example, from the data set, $x_1^{\{3, 4, 5, 6\}}$ is fourth order pattern, attributes and its values are [eggs=1, aquatic=0, backbone=0, tail=0]

Where attribute index set $^{[3, 4, 5, 6]}$ is referred to the attributes {eggs, aquatic, backbone, tail};

By the same token, $x_2^{[7, 8, 9, 10]}$ represents the pattern [milk=0, air bone=1, breaths=1, fins=0]

If we combine both x_1, x_2 as x_3, it will be $x_3^{[3, 4, 5, 6, 7, 8, 9, 10]}$ represents eggs=1, aquatic=0, backbone=0, tail=0, milk=0, airbone=1, breathens=1, fins=0.

->Pattern x_1 of attributes indexes [3, 4, 5, 6] are the same for the sample names.

Animal	Egg	Aqu	Bac	Tai
Gnat	1	0	0	0
Honeybee	1	0	0	0
Housefly	1	0	0	0
Ladybird	1	0	0	0
Moth	1	0	0	0
Wasp	1	0	0	0
Clam	1	0	0	0
Flea	1	0	0	0
Slug	1	0	0	0
Termite	1	0	0	0
Worm	1	0	0	0

Table3. Cluster P_1

From above table the attribute values for all animals are same .Thus we can consider it as a cluster
->Pattern x_2 with attribute values [7, 8, 9, and 10] are the same for the sample names.

Animal	Mil	Air	Bre	Fin
Chicken	0	1	1	0
Crow	0	1	1	0
Dove	0	1	1	0
Duck	0	1	1	0
Haming	0	1	1	0
Skua	0	1	1	0
Sparrow	0	1	1	0
Swan	0	1	1	0
Vulture	0	1	1	0
Wren	0	1	1	0
Gnat	0	1	1	0
Honeyb	0	1	1	0
Housefl	0	1	1	0
Ladybir	0	1	1	0
Moth	0	1	1	0
Wasp	0	1	1	0

Table.4 ClusterP_2

The pattern induced data clusters defined above are constant clusters .That is each cluster attributes has only one value. Since each pattern can induce a constant cluster, the no of constant clusters is overwhelming. To reduce number, it is desirable to merge clusters Let I (i) and I (j) be two data clusters induced by patterns and respectively. The merged data cluster of I (i) and I (j) is the union of their matched samples and matched attributes. Thus from the both obtained clusters P_1 and P_2 we can have another new cluster P_{12}. Clusters P_1 (Table.3) and P_2 (Table.4) we can have another cluster P_{12} (Table.5).

When two data clusters are merged, the corresponding patterns including them are simultaneously clustered shown in table.5

Animal	Egg	Aqu	Bac	Tai	Mil	Air	Bre	Fin
Gnat	1	0	0	0	0	1	1	0
Honeybee	1	0	0	0	0	1	1	0
Housefly	1	0	0	0	0	1	1	0
Ladybird	1	0	0	0	0	1	1	0
Moth	1	0	0	0	0	1	1	0
Wasp	1	0	0	0	0	1	1	0

Table.5 Cluster P_{12} from ClusterP_1 (Table.3) and cluster P_2 (Table.4)

Each group can be represented as a sample space S_n, n=1-k in Euclidean space. each group contains the random number of patterns. As shown in figure 2 all the data points ate represented in Euclidean space. Using x, y coordinates each pattern can be plotted as a point in Euclidian space. To find the distance between each point in Euclidian space we can depend on attribute values of each pattern. Divide the represented points into number of required clusters in Euclidian space as shown in figure 3.

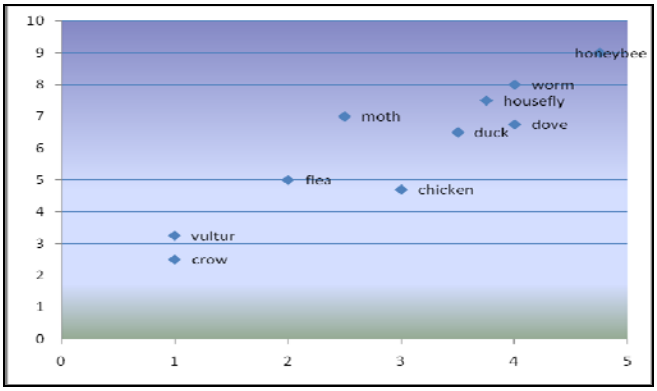


Fig.2 Representation of data set

Thus we have to choose k number of patterns as initial number of centriods and find the distance based on the attribute values. Let us make the relation by considering the attributes explicitly, they are milk, air bone, , fins and breaths. Thus we will get the cluster as shown in the figure.4.

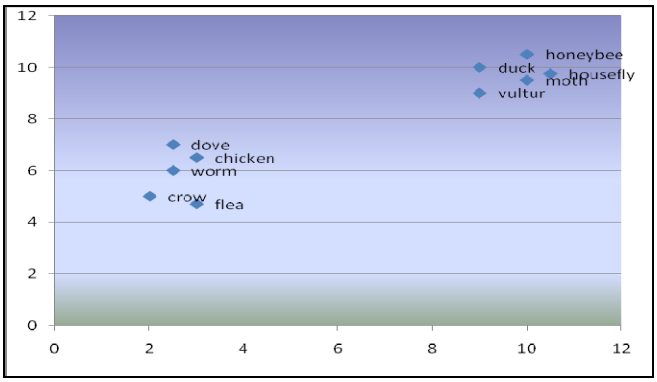


Fig.3 Dividing data into required no. of clusters

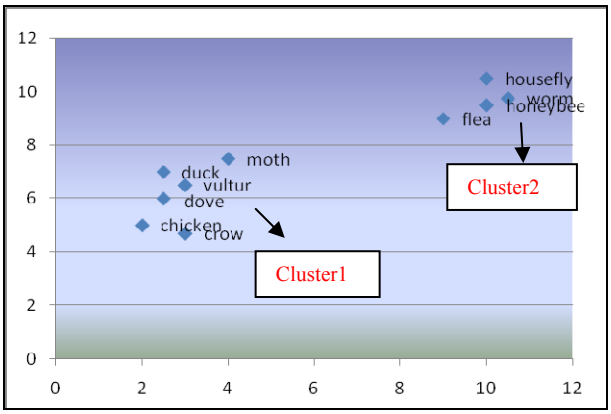


Fig.4 Cluster formed based on attributes [milk, air, bre, fin]

COMPARISON: In existing system, when d_D [2] is used, it takes 5 iterations to cluster 25 patterns. It will take more than 20 iterations to cluster 200 patterns. In many real-world data sets, the number of patterns produced by PD is largely in the thousand magnitudes. Thus, with the existing system the number of iterations will be more for complex data sets. Using the proposed algorithm, number of iteratios is reduced to 3 to cluster 25 patterns. Thus to cluster 200 patterns, it will take 12 iterations.

5 .CONCLUSION AND FUTURE SCOPE

This paper has proposed a method for clustering patterns and their associated data. The effectiveness of the above divide-and-conquer approach lies in the proposed clustering method. It is referred to as "**Simultaneous pattern and data clustering using modified K-Means algorithm**". One important property of the proposed method is that each pattern cluster is explicitly associated with a corresponding data cluster. To effectively cluster patterns and their associated data, several distance measures are used. Pattern pruning can be used before pattern clustering is the scope of our work.

REFERENCES

[1] Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 487-499, 1994.

[2] A.K.C. Wong, Fellow, IEEE, and Gary C.L.Li "Simultaneous pattern and data clustering for pattern cluster analysis" IEEE Trans. Knowledge and Data Eng., vol.20, no. 7, pp. 911-923, JULY 2008.

[3] S. Brin, R. Motwani, and R. Silverstein, "Beyond Market Basket:Generalizing Association Rules to Correlations," Proc. ACM-1997

[4] A.K.C. Wong and Y. Wang, "High Order Pattern Discovery from Discrete-Valued Data," IEEE Trans. Knowledge and Data Eng., vol. 9, no. 6, pp. 877-893, Nov./Dec. 1997.

[5] A.K.C. Wong and Y. Wang, "Pattern Discovery: A Data Driven Approach to Decision Support," IEEE Trans. Systems, Man, Cybernetics Part C, vol. 33, no. 1, pp. 114-124, 2003

[6] T.Chau and A.K.C. Wong, "Pattern Discovery by Residual Analysis and Recursive Partitioning," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 833-852, Nov./Dec. 1999.

[7] W.H. Au, K.C.C. Chan, A.K.C. Wong, and Y. Wang, "Attribute Clustering for Grouping, Selection and Classification of Gene Expression Data," ACM/IEEE Trans. Computational Biology and Bioinformatics, vol. 2, no. 2, pp. 83-101, Apr.-June 2005.

[8] Data mining concepts – Kamber.

[9] Data Clustering: A.K. JAIN Michigan State University, M.N. MURTY Indian Institute of Science AND P.J. FLYNN The Ohio State University.

[10] Arun K Pujari,"Data Mining Techniques", Universities press, 2001 edition.

[11] Jiawei Han and Micheline Kamber, "DATA MINING Concepts and Techniques" Elsevier Publishers, 2001 edition.

[12] S N Sivanandam and S Sumathi "Data Mining Concepts, Tasks and Techniques", Thomson Publishers, 2006 edition**.**

[13] P. Vijaya, M N Murthy and D K Subramanian. Leaders-sub leaders: An efficient hierarchical clustering algorithm for large data sets. Pattern Recognition Letters 25 (2004) 505-513.

[14] J.Han, Data Mining: Concepts and Techniques. Morgan Kautimann, 2001

[15] P.M. Murph and D.W.Aha, UCI Repository of Machine Learning Databases, Dept, Information and Computer Science, University of California, 1987.

[16] A. Silberschatz and A. Tuzhilin, "What Makes Patterns Interesting in Knowledge Discovery Systems" IEEE Irans. Knowledge and Data Engg. vol.8, no. 6, pp. 970-974, Dec. 1996.

[17] Modified K-means clustering algorithm Wei Li Institute of Operational Research & Cybernetics Hang Zhou Dianzi University Hang Zhou, 310018, China weili@hdu.edu.cn.