RESEARCH ARTICLE OPEN ACCESS

A LION OPTIMIZATION BASED K-PROTOTYPE CLUSTERING ALGORITHM FOR MIXED DATA

Mr.C.Mani M.C.A.,M.Phil.,M.E., [1] ,C. Mehala[2]

[1]Associate Professor, Department of Computer Science and Engineering,
Nandha Engineering College (Autonomus),Erode,Tamilnadu,India.

[2]Final MCA,Department of Computer Application,
Nandha Engineering College(Autonomus),Erode,Tamilnadu,India.

Email: [1]cmanimca@gmail.com,[2]mehalachinnasamy444@gmail.com

Abstract. Data Mining is used to gather information from huge set of data. Clustering is a grouping task for a set of objects. Clustering algorithms are divided by several types including hierarchical clustering algorithms,partitioning clustering and density based. The partitioning clustering includes K-Means clustering, K-Modes Clustering and CLARA algorithm. The K-Means clustering is only used for numeric data which has original optima. The K-Modes extends to the K-Means when the sphere is categorical. One of the most important algorithms for clustering heterogeneous type of data is the K- Prototype algorithm. This algorithm is veritably salutary for clustering large data sets. One of the simple optimization methods is Lion Optimization, that could be applied effectively for enhancing clustering results. It's useful for handling mixed data set. This leads a good optimization to calculate the centroid with K- Prototype clustering method. To overcome the problem in this clustering, Lion optimization Algorithm can be used. The proposed algorithm is enforced on standard standard dataset taken from UCI Machine Learning Repository. The Lion Optimization grounded K-Prototype clustering algorithm yields a better result when compared with the K- Prototype clustering.

Keywords: Kmeans Clustering, Lion Optimization, Data Mining, Machine Learning.

I. INTRODUCTION

Data mining is the analysis step in “ knowledge discovery in databases” or KDD. It's the process of sorting through large data sets to identify patterns and establish connections to break problems through data analysis. The data mining task issemi-automatic/ automatic analysis of data with large amounts for rooting preliminarily unknown, intriguing patterns similar as groups of data records (in cluster analysis), dependences (in association rule mining and successional pattern mining) and unusual records (in anomaly discovery).

There are two orders of data mining are Descriptive The descriptive orders include bracket, retrogression, time series. The order includes clustering, summarization, association rules, sequence discovery.

Data mining is generally classified as association, bracket, clustering, and vaticination (4). Within data mining, bracket/ vaticination are two kinds of data analysis used to wring models to describe Essential data module or to anticipate unborn data trends. The bracket system has two corridor the first part is learning practice, in which training data will be anatomized. The learned type or classifier shall be characterizing in the shape of bracket regulations.

The other position of bracket practice, in which test information to calculate roughly the fineness of bracket style or classifier. However, the regulations can be useful to bracket of new data, If the fineness is respectable. In fact, bracket system is supervised literacy, which is class position or analysis target is formerly known. As a result, the bracket form which is represented through rules structures will be constructed in the bracket system.

In this case, the created model will be representing the precious information and is use for forthcoming planning. Bracket is one kind of logical modeling. Further particularly, bracket is a conception for conveying rearmost objects to predefined type or classes from a collection of labeled records, construct the prototype similar as a decision trees and estimates markers for forthcoming not labeled records. Colorful bracket ways are KNN, K-Prototype; K- Captain are analysis and handed a comprehensive assessment for different bracket approaches in data mining.

CLUSTERING

Clustering is anun-supervised literacy. A cluster is a collection of objects which are analogous between them and are different to the objects belonging to other clusters. Clustering is also used to reduce the dimensionality of the

data when you're dealing with a riotous number of variables. The thing of clustering is to discover both the thick and the meager regions in a dataset. There are two types of clustering a) Hierarchical Clustering and
b) Partitional Clustering.

HIERARCHICAL CLUSTERING

The hierarchical clustering is an algorithm that groups analogous objects into groups. This scale of clusters is represented as a structure of a tree. There are two types of hierarchical clustering, Divisive and Agglomerative. In top-down or divisive clustering system we assign all of the compliances to a single cluster and also partition the cluster to two least analogous cluster. In bottom-up or agglomerative clustering system we assign each observation to its own cluster.

II. LITERATURE REVIEW

Zhexue Huang (1997), anatomized the K- Prototype clustering algorithms for mixed data similar as numeric and categorical data. The K-Means grounded styles have the effectiveness of the large datasets and it has limited numeric value to be estimated. In the exploration they've introduced a K- Prototypes algorithm grounded on the K-Means partitions to removes the numeric data limitation.

A system was developed to stoutly modernize the K-Prototype in order to maximize the intra cluster similarity of objects. The decision tree induction algorithm is used for creating rules for clusters and to understand and identify intriguing clusters.

Zhexue Huang (1998), have proposed a two algorithms which extend the K-Means algorithm to categorical disciplines and disciplines with mixed numeric and categorical values. The K-Modes algorithm uses a simple matching diversity measure to deal with categorical objects, replaces the means of clusters with modes, and utilizes a frequence-grounded method to modernize modes in clustering process to reduce clustering cost function.

With these extensions, K-Modes algorithm enables the clustering of categorical data in a fashion analogous to K-Means. They've concentrated on the specialized issues of extending the K-Means algorithm to cluster data with categorical values. Although they've demonstrated that the two new algorithms work well on two known data sets. It has to admit that this substantially rebounded from a priori knowledge to the data sets. In practical operations similar a priori knowledge is infrequently available. Thus, using the two algorithms is to break the particular data mining process.

Ming-Yi Shih et al (2010), has proposed a two step system for clustering mixed categorical and numeric data. Clustering algorithm work effectively with pure numeric data or pure categorical data. But it has been work inadequately with the mixed data similar as numeric and categorical data. The two step clustering was used for the diversity measures to deal with both categorical and numerical data. A two step system has been introduced for integrating hierarchal and partitioning clustering algorithm for the weakness of K-Means algorithm.

They've proposed a new approach as a single clustering algorithm to explore the relationship among the categorical values and numeric values. The categorical values have converted into numerical values and the numeric values were applied for the data sets.

Jinchoa Ji et al (2012), anatomized a fuzzy K-Prototype clustering algorithm for mixed numeric and categorical data. It has combined mean and fuzzy centroid for representing the prototype of the cluster. It have been employed a new measure which is grounded on theco-occurrence of values to estimate the diversity between data objects and prototypes of cluster.

The fuzzy c-mean clustering algorithm has been proposed to cluster these types of the data. The proposed algorithms were substantially used for high clustering delicacy, which have been demonstrated by experimental results. The performance grounded criterions were used to determine the optimal value for the fuzzy measure.

Li Xinwu (2012), proposed a new textbook clustering algorithm grounded on the bettered K-Means. A new textbook clustering were presented grounded on K-Means and Self-Organizing Model (SOM). The textbook have preprocessed to satisfy success process demand which have been bettered selection of original cluster center and cluster seed selection styles.

The K-Means improves the insufficiency of K-Means algorithm but the original cluster center is more sensitive and for the insulated point textbook. The advantages of K-Means and SOM were combined to a new model to the cluster textbook in the paper. The experimental results have shown that clustering combination algorithm which not only maintains the tone – organizing features of SOM network, but also makes up the disadvantages of SOM network’s overlong confluence duration. The bad clustering goods were caused by the shy selection of the K-Means algorithm’s original cluster center.

WuSen. et al (2013), has proposed a K- Prototype clustering algorithm for deficient datasets with mixed numeric and categorical attributes. The traditional K-Prototype algorithm is well clued in clustering data with mixed numeric and categorical attributes, while the completed data are limited. To handle deficient dataset with missing values, an advanced K- Prototype algorithm were proposed, which employs a new diversity measure for deficient dataset with mixed numeric and categorical attributes. A new approach was used to elect K objects as the original prototypes grounded on the nearest neighbors. To illustrate the delicacy of the established algorithm, traditional K- Prototype algorithm and K- Prototype employing the new diversity measure were compared to the bettered K- Prototype algorithm. The new diversity measure calculation takes into account missing data, with no need to attribute missing data with means or modes before clustering, which decreases an estimation that might beget some error.

Izhar Ahmad et al (2014), anatomized as K-Means and K-Prototype as performance analysis. The system design approach has been presented for the K-Means and K-Prototype performance analysis. The system armature in the exploration have presented an detail discussion of the K-Means and K-Prototype algorithm to recommend effective algorithm for outlier discovery and other issues which are related to the database clustering.

The original algorithm of the K-Means and K-Prototype algorithm does always the guarantee the delicacy of final clusters which are grounded on the selection of original centroids. The proposed system armature have use the complete unified result for the K-Means and K-Prototype algorithms of performance analysis. The analysis have shown that the proposed system armature procedures more clusters in lower calculation time as compared to the standard K-Means and K-Prototype algorithm.

Tang Zhe et al (2014), have proposed K-Means Clustering Algorithm system grounded on Scuffled frog springing algorithm to resolve the problems of the traditional K-Means Clustering algorithm similar as arbitrary opting of original clustering centers, the low effectiveness of clustering, low in the real. They've proposed a new K-Means clustering algorithm grounded on Scuffled frog springing algorithm. According to the variation of the frog's finess friction were used K-Means algorithm, it have been the advantages in the global hunt capability and confluence speed.

The experimental results have shown that the proposed algorithm have advanced delicacy. An algorithm grounded on SFLA K-Means clustering algorithm, on the base of K-Means clustering algorithm have overcome the delicacy of the unstable because K-Means sensitive to the original cluster centers and lower delicacy rate and other issues.

K.Arun Prabha et al (2015), discussed an alternate variant of double a) Flyspeck Swarm Optimization and b) K- Prototype algorithms for reaching global optimal result to cluster the optimization problem. The relative analysis of K-Prototype and PSO proved that Flyspeck Mass predicated on K- Prototype algorithm provides better performance than the traditional K-Modes and K-Prototype algorithms.

Particle Swarm Optimization predicated K-Prototype Clustering algorithm though incorporation of the benefit of it with K- Prototype algorithm, for reaching the global optimum cluster result. It is proved that exuction performance of the new algorithm was superior to the conventional K-Modes and K- Prototype algorithms performance.

Preeti Arora et al (2015), analyzed K-Means and K-Medoids for Big Data. In the paper the two K-Means and K-Medoids were estimated on dataset trade of Boat. The input of the algorithms were erratically distributed data points and predicated on their similarity clusters are generated.

The comparison results have proved that time taken in selection of cluster head and space complexity for lapping the cluster is better in K-Medoids than K-Means. It is also shown that K-Medoids are better in the entire aspects analogous as time of execution, on sensitive to outliers and noise reduction.

Izhar Ahmad et al (2014) said that the evaluation of computer and information technologies had changed the way users used to communicate as well as perform their execution tasks. Data booby-trapping algorithm is employed for converting the data in to knowledge information that can be used in future for performing different tasks. In this disquisition, they have presented a system design approach for the K-Mean and K-Prototype Algorithms Performance Analysis.

The system architecture in this disquisition is presents a detail discussion of the k- means and k-prototype to recommend effective algorithm for outlier discovery and other issues relating to the database clustering. The system design approach is predicated on the open source technologies. The verification and evidence of the system is predicated on the simulation.

III. PROPOSED METHODOLOGY

The main objective in this work is to optimize the K-Prototype clustering using Lion Optimization Algorithm. A numerical dataset needs to have an approximate previous knowledge of enrolled class id to prognosticate their performance in future values. The difficulties in dealing with high-dimensional data are universal and abundant. Still, not all marvels that arise are inescapably mischievous to clustering ways.

The nomad, which is the tendency of some data points in high-dimensional data sets to do much more constantly in k-nearest-neighbor lists of other points than the remaining points from set, are in fact be used to cluster. In this design concentrated on exploring the implicit value of using mecca points in clustering by designing nomad-apprehensive clustering algorithms and testing them in a high-dimensional environment.

The nomad is a good measure of point centrality within a high-dimensional data cluster and that major capitals can be used effectively as cluster prototypes. Centroids and medoids in K- means duplications tend to nverge to locales close to high-vagabond points, which implies that using rambler rather of either of these could actually speed up the confluence of the algorithms, leading straight to the promising regions in the data space.

A simple way to employ rambler for clustering is to use them as one would typically use centroids. Indeed though points with loftiest rambler scores are without mistrustfulness the high campaigners for cluster centers, there's no need to disregard the information about rambler scores of other points in the data.

The system is defined with the following specific objects.

- A Lion Optimization Algorithm is induced to recitify the K- Prototype algorithm's problem.
- 5 datasets are taken from UCI repository; these are given to Lion Optimization and K- Prototype algorithms. The outcomes are estimated and validity measure like Rand indication, F-Measure, Jaccard indicator and Entropy are used.

The clustering system is used to identify academically at threat scholars and classify the scholars consequently, the K-Means system is used then. MapReduce is one of the major factors for distributed data processing. MapReduce programming model consists of two separate and distinct tasks.

The first is the chart job, which takes a set of data and converts it into another set of data, where individual rudiments are broken down into tuples (key/ value dyads). The reduce job takes the affair from a chart as input and combines those data tuples into a lower set of tuples.

The chart function shows the procedure of assigning each sample to the closest center while the reduce function performs the procedure of streamlining the new centers. In order to drop the cost of network communication, a combiner function is developed to deal with partial combination of the intermediate values with the same key within the same chart task. To employ capitals for clustering is to use them as one would typically use centroids.

The Lion Optimization Algorithm (LOA) is a population based meta-heuristic algorithm which randomly generate the population over the solution. In the algorithm every single solution is called “Lion”. In an Nvar dimensional optimization problem. A lion is represented as

Lion=[x1, x2, x3,……, xNvar]
(5)

The cost function of each lion is computed by

Fitness value of Lion=f (Lion) =f(x1, x2, x3,…xNvar) (6)

A dummy prey (PREY) is considered as the center of hunters which is calculated by

PREY=∑hunters(x1, x2, x3,……xNvar)/number of hunters (7)

During hunting, hunters are selected randomly and the selected hunter attack with the dummy prey which the selected lion belongs to which group. If the prey escapes from the hunter, the new position of the lion is calculated by

PREY’=PREY+rand (0, 1)*PI*(PREY-Hunter)
(8)

where PREY is current position of prey, Hunter is new position hunter who attack to prey and PI is the percentage of improvement in fitness of hunter. The new position of the hunter is calculated using

Hunter’={ rand((2*PREY-Hunter),PREY),(2*PREY-Hunter)<PREY
rand((2*PREY-Hunter),PREY),(2*PREY-Hunter)<PREY

The new position of the Female lion is calculated by

Female Lion’=Female Lion+2D *rand(0,1){R1}+U(-1,1)*tan(θ)*D*{R2}{R1}.{R2}=0 (10)

Table 3.1 LION OPTIMIZATION ALGORITHM

IV. FINDINGS

- K-Means, K-Prototype and Lion Optimization grounded K-Prototype clustering algorithm’s performance are measured in terms of external validity measures like F-Measure, Rand Index, Jaccard Index and Entropy.
- The external validity measures shows the quality of clusters by comparing the clustering results.
- All these measures have a value between zero and one.
- In case of Rand Index, Jaccard Index and F-Measure, the value one indicates that the data clusters are exactly same and so increase in the values of these measures proves the better performance.
- The trial analysis is performed with text datasets.

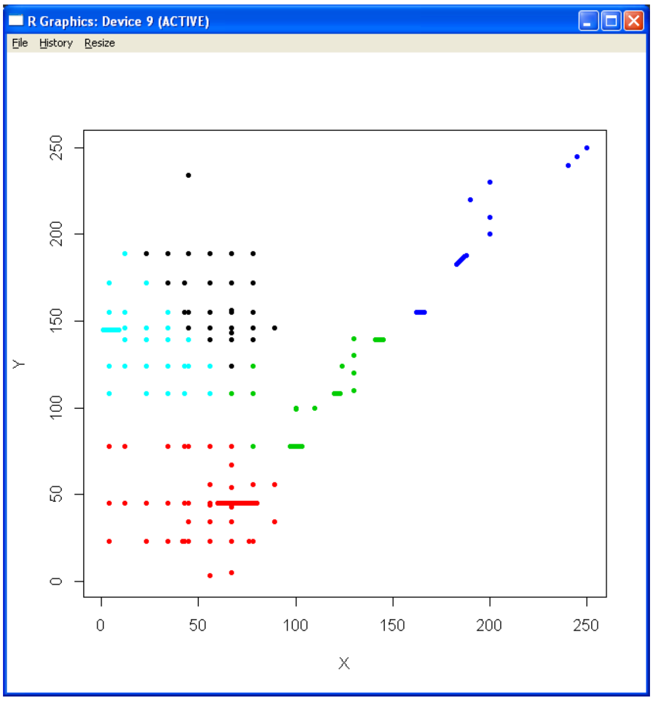
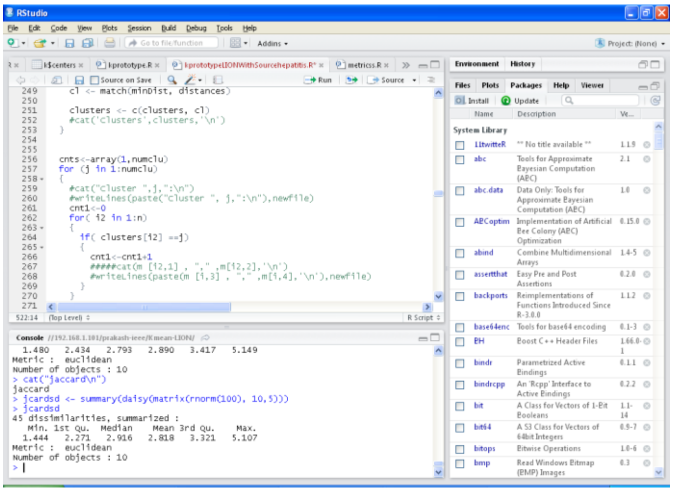


Fig 4.1 K-LION MULTI LEVEL CLUSTERING



```
    cl <- match(minDist, distances)

    clusters <- c(clusters, cl)
    #cat('Clusters',clusters,'\n')
  }


  cnts<-array(1,numclu)
  for (j in 1:numclu)
  {
    #cat("Cluster ",j,":\n")
    #writeLines(paste("Cluster ", j,":\n"),newfile)
    cnt1<-0
    for( i2 in 1:n)
    {
      if( clusters[i2] ==j)
      {
        cnt1<-cnt1+1
        #####cat(m [i2,1] , "," ,m[i2,2],'\n')
        #writeLines(paste(m [i,3] , "," ,m[i,4],'\n'),newfile)
      }
    }
```

```
  1.480   2.434   2.793   2.890   3.417   5.149
Metric :  euclidean
Number of objects : 10
> cat("jaccard\n")
jaccard
> jcardsd <- summary(daisy(matrix(rnorm(100), 10,5)))
> jcardsd
45 dissimilarities, summarized :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.444   2.271   2.916   2.818   3.321   5.107
Metric :  euclidean
Number of objects : 10
> |
```

	Name	Description	Ve...
System Library			
□	1ltwitteR	** No title available **	1.1.9
□	abc	Tools for Approximate Bayesian Computation (ABC)	2.1
□	abc.data	Data Only: Tools for Approximate Bayesian Computation (ABC)	1.0
□	ABCoptim	Implementation of Artificial Bee Colony (ABC) Optimization	0.15.0
□	abind	Combine Multidimensional Arrays	1.4-5
□	assertthat	Easy Pre and Post Assertions	0.2.0
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.2
□	base64enc	Tools for base64 encoding	0.1-3
□	BH	Boost C++ Header Files	1.66.0-1
□	bindr	Parametrized Active Bindings	0.1.1
□	bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2.2
□	bit	A Class for Vectors of 1-Bit Booleans	1.1-14
□	bit64	A S3 Class for Vectors of 64bit Integers	0.9-7
□	bitops	Bitwise Operations	1.0-6
□	bmp	Read Windows Bitmap (BMP) Images	0.3

Fig 4.2 JACCARD SIMILARITY

V. CONCLUSION

This paper proposed Lion Optimization grounded K-Prototype Clustering algorithm by incorporating the benefit of Lion Optimization algorithm with the being K-Prototype algorithm, to reach the global optimum cluster result. The proposed algorithm has been tested on the five standard datasets which include both numeric and categorical attributes. It is proved that the execution performance of new algorithms are much better than conventional K- Means and K- Prototype clustering algorithms' performance.

- In future, applicable optimization algorithm will be applied for tuning of parameter to produce superior quality clusters.
- The global cluster results can further be bettered by setting alternate values for the parameters of Lion Optimization Algorithm.

REFERENCES

[1] Arun Prabha .K, N. Karthi Keyani Visalakshi,"Particle Swarm Optimization based K-Prototype Clustering Algorithm" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, ISSN: 2278-8727, Vol 17, pp. 56-62, [2015].

[2] Arun K Pujari,"Data Mining Techniques", Third Edison ISBN: 978-81-7371-884-7[2013].

[3] Fengmei W and H. Lixia, "A Missing Data Imputation Method Based on Neighbor Rules", Computer Engineering, vol. 38, no. 21, [2012].

[4] Gong Jing, Li Anming, "An Implementation of Clustering Algorithm Based on K-means", Journal of Hunan University of Technology, vol. 22, pp. 52–54, [2008].

[5] Izhar Ahmad,"K-Mean and K-Prototype Algorithms Performance Analysis", American Review of Mathematics and Statistics, ISSN 2374-2348, Vol. 2, pp. 95-109, [2014].

[6] Jinchao Ji, Wei Pang, Chunguang Zhou, Xiao Han,Zhe Wang,"A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data", Elsevier , ISSN 0950-7051,Vol. 30,[2012].

[7] Li Xinwu,"A New Text Clustering Algorithm Based on Improved K-Means", Journal of Software, Vol. 7, doi:10.4304/jsw.7.1.95-101, [2012].

[8] Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai," A Two-Step Method for Clustering Mixed Categroical and Numeric Data", Tamkang Journal of Science and Engineering, Vol. 13, pp. 11-19, [2010].

[9] Maziar Yazdani, Fariborz Jolai,"Lion Optimization Algorithm (LOA): A Nature-Inspired Metaheuristic Algorithm", Elsevier, Journal of Computational Engineering, [2015].

[10] Madhuri R, M Ramakrishna Murty, JVR Murthy, PVGD Prasad Reddy, and Suresh C Satapathy, "Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms," ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol II, pp. 137-144, [2014].

[11] Meng Ying, Luo Ke, et al,"K-medoids clustering algorithm method based on ant colony Algorithm", J.Computer Engineering and Applications, 48(16):136-139, [2012].

[12] Navneet, Nasib Singh Gill,"A Novel Algorithm for Big Data Classification Based on Lion Optimization", Journal of Theoretical and Applied Information Technology, ISSN: 1992-8645, E-ISSN: 1817-3195, Vol.95, [2017].

[13] Preeti Arora, Dr. Deepali , Shipra Varshney ”Analysis of K-Means and K-Medoids Algorithm For Big Data” Elsevier,Procedia Computer Science 78, pp. 507 – 512, [2016].

[14] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, “Introduction to Data Mining”, Pearson, ISBN: 978-93-82291-49-7 [2015].

[15] T.V Suresh Kumar, B.Eswara Reddy, Jagadish S Kallimani,”Data Mining Principles & Applications”, Elsevier, ISBN: 978-93-82291-49-7 [2013].

[16] Tang Zhe, Luo Keb,”K-means Clustering Algorithm Method Based on Shuffled Frog Leaping Algorithm”, Advanced Materials Research, Vol. 989-994, pp 2245-2249,[2014]

[17] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets.html.

[18] Wu Sen, Chen Hong and Feng Xiaodong ,”Clustering Algorithm for Incomplete Data Sets with Mixed Numeric and Categorical Attributes”, International Journal of Database Theory and Application, ISSN: 2005-4270 IJDTA ,Vol.6, pp.95-104,[2013].

[19] Zhexue Huang,”Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values”, Data Mining and Knowledge Discovery 2, pp. 283–304, [1998].

[20] Zhexue Huang,”Clustering Large Data Sets with Mixed Numeric and Categorical Values”, CSIRO Mathematical and Information Sciences, Conference, pp-21-34, [1997].

[21] C.-H. Cheng and Y.-S. Chen, ‘‘Classifying the segmentation of customer value via RFM model and RS theory,’’ Expert Syst. Appl., vol. 36, no. 3, pp. 4176–4184, Apr. 2009.

[22] C.-C.-H. Chan, C.-B. Cheng, and W.-C. Hsien, ‘‘Pricing and promotion strategies of an online shop based on customer segmentation and multiple objective decision making,’’ Expert Syst. Appl., vol. 38, no. 12, pp. 14585–14591, Nov. 2011.

[23] R.-S. Wu and P.-H. Chou, ‘‘Customer segmentation of multiple category data in e-commerce using a soft-clustering approach,’’ Electron. Commerce Res. Appl., vol. 10, no. 3, pp. 331–341, May 2011.