Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 6, July, 2021: 8507 - 8515

Research Article

Data Clustering Of Numerical And Categorical Datasets Using Harmony Search Based Ensemble Technique

Dr. Muhammed Basheer[1], Ms. Jaya Khatri[2], Ms. Preeta Rajiv Sivaraman[3], Smt.Z.Sunitha Bai[4]

Abstract

Clustering is a common method for finding patterns in underlying data in data mining applications. The majority of conventional clustering methods are restricted to datasets with numeric or categorical characteristics. In real-world data mining applications, however, datasets containing various kinds of characteristics are frequent. To address this issue, we offer a new divide-and-conquer strategy in this article. To begin, the original mixed dataset is split into two sub-datasets: pure category and pure numeric. Then, to generate matching clusters, existing well-established clustering algorithms intended for various kinds of datasets are used. The superiority of our method is shown by comparisons with existing clustering algorithms on real-world datasets. Clustering is a well-known data mining method for pattern detection and retrieval of information. The data in the first clustering dataset may be categorised or numerical. Each kind of data has its own method for clustering. The k-means method for clustering numeric datasets and the k-modes technique for categorical datasets are proposed in this area. The transformation of categorical characteristics into numeric measurements and direct application of the k-means algorithm instead of the k-modes method is one of the major issues in achieving the clustering process on categorical values. In this article, it is suggested to test a method based on the preceding problem, which involves converting categorical data into numeric values by utilising the relative frequency of each modality in the characteristics.

Keywords: Clustering, Data Mining, Categorical Data, *k*-means, categorical datasets.

1. Introduction

Clustering divides data into groups with the goal of maximising intra-cluster similarity while minimising inter-cluster similarity. In various areas, such as pattern recognition, customer segmentation, similarity search, and trend analysis, the clustering method has been widely explored. Because datasets with mixed kinds of characteristics are prevalent in real-world data mining applications, the ability to deal with datasets including both numeric and categorical variables is unquestionably essential [1]. Although numerous clustering methods have been

[1]Assistant Professor Department of Mathematics. IT University of Technology and Applied science Oman
[2]Assistant Professor , Asian School of Business, Noida. Email-ID: jayakhatri17@gmail.com
[3]Assistant Professor ,Asian School of Business, Noida. Email ID - preetasiva@gmail.com
[4]Assistant professor, Department of Computer science and engineering, RVR&JC College of Engineering, Guntur

developed to date, the majority of them are based on the premise that all characteristics are either numeric or categorical. The majority of prior clustering methods have focused on numerical data, which has intrinsic geometric characteristics that may be used to construct distance functions between data points. However, most of the data in the databases was categorical, with attribute values that couldn't be arranged organically as number values. Shape is an example of a categorical attribute, with values such as circle, rectangle, and ellipse [2]. Attempts to create criterion functions for mixed data have been unsuccessful due to the variations in the properties of these two types of data. To begin, the original mixed dataset is split into two sub-datasets: pure category and pure numeric. Then, to generate matching clusters, existing well-established clustering algorithms intended for various kinds of datasets are used. Finally, the clustering results from the categorical and numeric datasets are merged to form a categorical dataset, which is then used to generate the final output using the categorical data clustering method. Because of its capacity to tackle nonlinear and complicated optimization problems, evolutionary algorithms (EAs) inspired by Darwinian theory of evolution have grown in popularity over the past decade. EAs, unlike traditional numerical optimization techniques, are population-based metaheuristic algorithms that just need the objective function values and do not require characteristics like differentiability or continuity [3]. The performance of EAs, on the other hand, is influenced by encoding methods, evolutionary operators, and parameter settings such as population size, mutation scale factor, and crossover rate. Furthermore, proper parameter selection is problem-dependent and necessitates a time-consuming trial-and-error parameter tuning procedure. If the optimization is needed in an automated environment or if the user has no expertise with the fine art of control parameter tuning, trial-and-error based parameter selection is unsuccessful. Different parameter adaption methods have been proposed to address this. Adaptive and selfadaptive methods are prominent among the many parameter adaptation strategies because they may change the parameter throughout the evolution with little or no involvement from the user [4]. In other words, parameter adaptation in adaptive and self-adaptive methods is dependent on input from the search process. The premise behind self-adaptive methods is that the most suitable parameter values create better offspring, who are more likely to survive and spread the superior parameter values. As a result, with self-adaptive techniques, the parameters are directly encoded into the people, and the encoded solutions develop alongside them. The rapid growth in the manufacture of information technology devices, as well as improvements in scientific data collecting techniques, has resulted in the establishment of ever-increasing data archives. Furthermore, conventional exploratory techniques have shown their inefficiency in dealing with large amounts of data in order to uncover novel discoveries. As a result, recently created knowledge-discovery systems should use novel and suitable machine learning techniques to investigate these massive structures and uncover previously concealed patterns. Clustering is the most popular knowledge-discovery method used in data mining for information retrieval and pattern identification. It refers to unsupervised learning with the goal of partitioning a dataset of N people embedded in d-dimensional space into K unique clusters without any previous knowledge of the cluster distribution. The results show that data points in the same cluster are more similar than data points in other clusters [5].

2. Literature Review In Clustering Categorical Datasets

Categorical Clustering Algorithms:

Although many suggestions have been made in the area of clustering categorical datasets, the k-mode and its variations are the most often used. It's a variant of the k-means method that

replaces the Euclidean distance with a simple matching dissimilarity function that's more suited to categorical values, and the means with the modes to find the most representative member in a cluster [6]. Furthermore, the modes are based on a frequency-based technique for updating the centroids in each iteration. Clustering mixed datasets with category and numeric variables is possible using the k-prototype method. The fuzzy k-modes algorithm and the fuzzy k-modes method with fuzzy centroids are two versions that have been suggested. The major disadvantage of utilising the simple dissimilarity matching distance is that it does not provide efficient results since simple matching often produces clusters with low intra-similarity. The authors demonstrated that the similarity between two categorical values may also be referred to as their co-occurrence according to a common value or set of values, which is the second method for clustering categorical data based on attribute co-occurrence. The ROCK algorithm is the most well-known example of this kind of algorithm. It uses the notion of links to quantify the similarity of categorical patterns, i.e. the similarity between any two categorical patterns is proportional to the number of common neighbours they have. As a result, the goal of this method is to combine the patterns into a group with a high number of connections. Since the simple matching distance metric is not a good measure because it results in poor intra-cluster similarity, the concept of relative frequency was used to define a new dissimilarity coefficient for the k-modes algorithm in which the frequency of categorical values in the current cluster was considered to calculate the dissimilarity between a data point and a cluster mode [7].

Although k-modes based algorithms, like k-means types algorithms, have demonstrated their efficacy in clustering large categorical datasets, they still have two major drawbacks: the inability to cover global information effectively, i.e. the provided solutions are only local optimal and a global solution is difficult to find, and the accuracy of the obtained results is sensitive to the number of k-modes. Furthermore, since the attribute values of categorical data are not continuous, the modes are more difficult to shift in iterative optimization procedures. Because the mode reflects the most common element in the examined modality, if two modalities have similar frequencies, only one will be kept and the other will be rejected, resulting in information loss [8].

In this progression, the optimization dispute is determined as follows:

$$Minimize\ g\ (y) \tag{1.1}$$

$$Subject\ to\ y_i \in\ Y_i, i\ = 1, 2, \ldots, N\ . \tag{1.2}$$

$where,$

$g\ (y)\ -\ An\ objective\ function;$

$y -\ Set\ of\ each\ \ decision\ variable\ \ y_i\ ;$

$Y_i - Set\ of\ possible\ range\ of\ \ values\ for\ each\ decision\ variable,$

$(i.e).,Y_i\ =\ \{\ y_i\ (1), y_i(2), \ldots, y_i\ (K)\}$

$for\ discrete\ decision\ variables$

$\left(\ y_i\ (1) <\ \ y_i\ (2) <\ \ldots\ <\ \ y_i\ (K)\right);$

$N -\ \ No.of\ decision\ variables\ (number\ of\ Musical\ instruments);$

$K - No.\ \ of\ favourable\ values\ for\ the\ discrete\ variables$

$(Pitch\ Range\ of\ \ each\ musical\ instruments).$

Harmony search matrix (HSM) is loaded up with however many haphazardly created arrangement vectors as the size of the harmony search (HS).

$$\begin{bmatrix} {y_1}^1 & {y_2}^1 & \cdots & {y_{N-1}}^1 & {y_N}^1 \\ {y_1}^2 & {y_2}^2 & \cdots & {y_{N-1}}^2 & {y_N}^2 \\ \vdots & \cdots & \cdots & \cdots & \cdots \\ {y_1}^{HSM-1} & {y_2}^{HSM-1} & \cdots & {y_{N-1}}^{HSM-1} & {Y_N}^{HSM-1} \\ {y_1}^{HSM} & {y_2}^{HSM} & \cdots & {y_{N-1}}^{HSM} & {y_N}^{HSM} \end{bmatrix} = \begin{bmatrix} g(y^1) \\ g(y^1) \\ \vdots \\ g(y^{HSM-1}) \\ g(y^{HSM}) \end{bmatrix} \quad (1.3)$$

Parameter Adaptation in Differential Evolution:

Although DE has lately received a lot of attention as a global optimizer for continuous spaces, the performance of the traditional DE algorithm is highly dependent on the mutation and crossover methods used, as well as the control settings used. The performance of DE becomes increasingly sensitive to the strategies and related parameter values as the issue complexity increases, and an incorrect choice may result in premature convergence, stagnation, or waste of computing resources. To put it another way, owing to the complicated interplay of control factors with DE performance, selecting suitable mutation and crossover methods and control settings necessitates considerable skill [9].

We are applying stochastic optimization population algorithm, the general differential evolution as given below:

$$y_i{}^{new} = y_i{}^{best} + F[(y_i{}^{r_1} - y_i{}^{r_2}) + (y_i{}^{r_3} - y_i{}^{r_4})\,]$$

$where,$

$r_1, r_2, r_3\ and\ r_4 - mututally\ independent\ variables$

$F - step\ size\ \ scale\ Factor$

Scaled differential vectors as for the conceivable individual sets adjust the property of the current area scene. It in this manner can furnish promising transformation bearings with flexible advance size and a harmony among nearby and worldwide inquiry.

Cluster Ensembles:

Cluster ensembles are a technique for combining several runs of various clustering methods to produce a common division of the original dataset, with the goal of consolidating findings from a portfolio of separate clustering outcomes. Although research on cluster ensembles is not as well known as research on multiple classifier or regression models, many independent studies have lately been conducted. On the basis of a hyper-graph model, we offer combiners for addressing the cluster ensemble issue as an optimization problem. The results of many separate runs of the same clustering algorithm are properly integrated in their technique to produce a data partition that is unaffected by initialization and overcomes the instabilities of clustering methods. The fusion process then begins using the clusters created by the combining phase and determines the optimum number of clusters based on certain specified parameters. To enhance clustering performance, the authors suggested a sequential combination technique. Their method first produces an initial result using global criteria-based clustering, then utilises local criteria-based information to enhance the original result using a probabilistic relaxation algorithm or linear additive model [10].

Cluster Ensemble: The Viewpoint of Categorical Data Clustering

In this part, we'll argue that the cluster ensemble issue may be recast as a categorical data clustering problem, and we'll show why current cluster combination techniques are ineffective in this context. Clustering attempts to find groupings in data sets and uncover interesting distributions and trends [11]. A particular clustering algorithm's output will typically be the assignment of data items in a

dataset to various groups. In other words, a unique cluster label will suffice to identify each data item. Data items with distinct cluster labels are regarded to be in separate clusters from the standpoint of clustering; if two objects are in the same cluster, they are completely similar, if not, they are fully dissimilar. As a result, it is clear that cluster labels cannot be given a natural ordering in the same way that real numbers can, implying that the clustering algorithm's output is categorical (or nominal). Cluster ensemble is a technique for combining several runs of various clustering algorithms to produce a common division of the original dataset, with the goal of consolidating findings from a portfolio of separate clustering outcomes. The following are some of the benefits of transforming the cluster ensemble issue into a categorical data clustering problem. For starters, several effective methods for grouping categorical data have recently been presented [12]. These methods can be fully used, and advancements in categorical data clustering research may help the cluster ensemble issue. Furthermore, categorical data clustering is a straightforward issue with a uniform framework for problem formalisation. We present a new divide-and-conquer method for clustering datasets with mixed kinds of characteristics. To begin, the original mixed dataset is split into two sub-datasets: pure category and pure numeric. Then, to generate matching clusters, existing well-established clustering algorithms intended for various kinds of datasets are used [13]. Finally, the categorical and numeric datasets' clustering results are merged to form a categorical dataset, on which the categorical data clustering method is used to get the final clusters. Now we'll concentrate on the last stage of our method. In reality, using current cluster ensemble techniques and the clustering findings from the categorical and numeric datasets, we can get the final clustering results. Existing cluster combination techniques, on the other hand, have their own set of restrictions and are not suitable for our issue. When conducting combination or categorical data clustering, the weights for the clustering output on categorical and numeric datasets [14]. Unfortunately, the method is intended to combine runs of clustering algorithms with the same number of clusters, and its ability to handle a weighted cluster ensemble issue is unclear. The issue with the sequential combination technique described in is the same as with the approach. Furthermore, their method is limited to combining the results of just two clustering algorithms. To address the cluster ensemble issue, the suggested combiners are based on a hyper-graph model. Their methods are simple to adapt to the weighted cluster ensemble issue; nevertheless, they all have significant computing costs, which are prohibitive in the context of access to huge data sets stored in secondary memory.

3. Cluster Ensemble Based Algorithm

Clustering mixed category and numeric data using a cluster ensemble-based algorithm framework. We'll start with a high-level review of the algorithm framework, then go on to the specifics and complexity findings in following parts. The cluster ensemble based algorithm framework and the processes required. To begin, the original mixed dataset is split into two sub-datasets: pure category and pure numeric. Then, to generate matching clusters, existing well-established clustering algorithms intended for various kinds of datasets are used. Finally, the categorical and numeric datasets' clustering results are merged to form a categorical dataset, on which the categorical data clustering method is used to get the final clusters. Because this method framework obtains clustering results from both categorical and numeric datasets, it is known as CEBMDC (Cluster Ensemble Based Mixed Data Clustering) Table 1.

Table.1. Set of Data Basic Informations.

DATA	# INST	# FEATURES	#CLASSES

CBIR	1654	195	7
CHART	500	70	5
ISOLET6	1540	718	5
SEGMENTATION	2410	29	6
EOS	2450	15	7
WINE	187	12	2

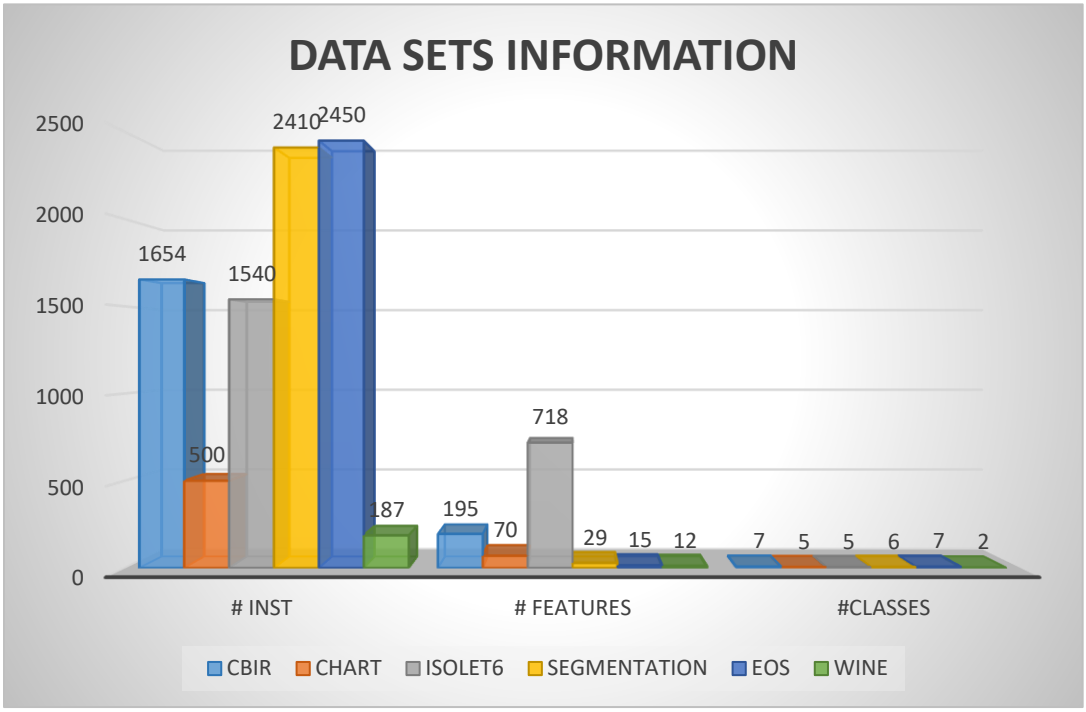

Fig.1. Bar diagram of the different data sets.

It ought to be noticed that each of the six informational indexes are named also, contain regulated class data Fig.1. The class names, notwithstanding, were just utilized in assessing the last grouping arrangements and not utilized at all during grouping or outfit choice.

Harmony Search Algorithm:

HS is a population-based metaheuristic optimization technique that is designed to mimic the process of musical improvisation, in which musicians improvise the pitch of their instruments in order to achieve a perfect state of harmony. A few characteristics of HS that distinguish it from other metaheuristics such as decomposition include the fact that it considers all existing solution vectors when creating a new vector, rather than just two as is the case with DE, and the fact that each decision variable in a solution vector is considered independently. It is important to note that the improvisation operators as well as memory consideration, pitch adjustment, and random consideration all play important roles in creating the optimal balance between exploitation and exploration while optimising an HS system. Pitch adjustment and random consideration are, in essence, the two most important components of achieving the necessary variety in high school. When using random consideration, the new vector's components are generated at random, and it has the same level of efficiency as other randomization techniques. The study of HS at random allows for the exploration of fresh regions in the search space that may not have been examined before by the researchers. Increased variety in HS is achieved via pitch adjustment, which adjusts

the components of a new vector's components within a certain bandwidth by adding or removing a small random amount from an existing component stored in HM. The pitch adjustment operator may also be regarded of as a device that aids in the promotion of the intensification of HS by controlling the probability of PAR occurring. The third HS operator, memory consideration, is responsible for the intensification of the HS algorithm's results. A high rate of harmonic acceptance implies that good historical/memory solutions are more likely to be selected or accepted than inferior ones. This may be compared to a certain degree of elitism in society. If the acceptance rate is too low, solutions will converge more slowly than they otherwise would. The HS technique has lately gained a great deal of attention from the scientific world, and it has shown to be an efficient tool for solving a wide range of optimization problems in the fields of engineering and computing. As a consequence of the increased interest in HS, the organization's performance has improved and evolved in response to the demands of the problems that have been dealt with.

Harmony Search Based Parameter Ensemble Adaptation for DE (HSPEADE):

As mentioned in the preceding section, various optimization issues need different mutation and crossover methods coupled with different parameter values to achieve optimum performance, depending on the nature of the problem (unimodal or multimodal) and available computing resources. Furthermore, various mutation and crossover strategies with different parameter settings may be preferable at different phases of evolution to solve a particular issue than a single set of strategies with unique parameter values as in the traditional DE to solve a specific problem. In response of these findings, an ensemble method was developed in which a pool of mutation and crossover strategies, as well as a pool of values for each related parameter, compete to generate successful offspring populations. DE population member is randomly given a mutation and crossover strategy, as well as parameter values from the relevant pools. The population members generate children utilising the methods and parameter values that have been allocated to them. If the created trial vector is capable of progressing to the next stage of evolution, the combination of methods and parameter values used to create the trail vector is saved. If a trial vector fails to join the next generation, the strategies and parameter values associated with that target vector are randomly reinitialized with equal probability from the respective pools or from the successful combinations saved.

To get optimum results using the ensemble method, the candidate pool of strategies and parameters should be limited to prevent the negative effects of ineffective strategies and parameters. In other words, the strategies and parameters in the various pools should have varied features so that they may show different performance characteristics at different phases of the development while dealing with a certain issue. Because the strategy and parameter pools of EPSDE are extremely restricted, most of the people in the pools may become outdated as the DE population evolves. As a result, it would be ideal if the strategy and parameter pools could change in tandem with the DE population. We propose an HS-based parameter ensemble adaptation for DE based on this rationale (HSPEADE). Algorithm depicts the general perspective of the proposed HSPEADE. The HM of the HS method is initialised with HMS number of randomly generated vectors once the DE population is initialised. The parameter combinations (F and CR values) related to the mutation and crossover methods employed are represented by the members of the HMare. Using the elements in the HM, the HS method generates a new parameter combination vector. During the evolution, each of the HMS + 1 parameter combinations is assessed by testing them on the DE population. The HM is modified as inHS algorithm after assessing all members of the HM and the newly produced parameter combination. Throughout the evolution phase of the DE algorithm, new

parameter combinations are generated and the HM is updated. The parameter combinations in HM should be varied throughout the early generations of the DE population evolution and should converge to the optimum combination towards the conclusion of the evolution to achieve optimal performance based on the ensemble method.

According to the harmony search based ensamble standard method algorithm as follows,

STEP 1: Randomly initialize a population of NP, D-dimensional Parameter vectors. set the generation number G=0.

STEP 2 : ***WHILE*** stopping criterion is not satisfied

DO

Mutation—Equation (2)

Crossover—Equation (3)

Selection—Equation (4)

Increment the generation count G = G + 1

STEP 3: ***END WHILE***

Algorithm.1: Standard DE Algorithm.

4. Experimental Setup and Results

The performance of the suggested parameter adaption method for DE is evaluated in this section. Below are the test problems, experimental settings, and comparison methods. To assess our method, we performed a thorough performance analysis. These trials and their outcomes are described in this section. We compared our algorithm's clustering performance to that of other methods using real-world datasets from the UCI Machine Learning Repository. Simultaneously, its characteristics are being investigated experimentally.

The calculation techniques, which are important for acquiring the gathering model, are portrayed. The gathering model is proposed to have the accompanying design.

$$P_{Ensamble}^{i} = \sum_{n}^{j=1} \alpha_j \times P_j^i$$

$where,$

$\alpha_j -$ weights of the learners.

$P_j^i - vector\ predicted\ \ model\ flows\ for\ i\ and\ j$

5. Conclusions

To address this issue, we offer a new divide-and-conquer strategy in this article. To begin, the original mixed dataset is split into two sub-sets: a pure category dataset and a pure numerical dataset. Then, to generate matching clusters, existing well-established clustering algorithms intended for various kinds of datasets may be used. Finally, the categorical and numeric datasets' clustering results are merged to form a categorical dataset, on which the categorical data clustering method is used to get the final clusters. The main contribution of this paper is to provide an algorithm framework for the mixed attributes clustering problem, in which existing clustering algorithms can be easily integrated, and the capabilities of various clustering algorithms as well as the characteristics of various datasets can be fully exploited. To get a better understanding of this approach, we will explore incorporating additional alternative clustering algorithms into the

algorithm framework in the future. Furthermore, the suggested divide-and-conquer method will be used to identifying cluster-based local outliers in big database environments with mixed type characteristics. Clustering categorical data is a difficult and time-consuming process that necessitates the development of a particular clustering method. The relative frequency of each modality in its characteristics is utilised to convert categorical measurements into numeric values in this article. The resultant dataset is then sent to the k-means algorithm. These results demonstrate the significant impact made by the usage of relative frequency.

Reference

[1] Z. He, X. Xu, S. Deng: "*Squeezer:* An Efficient Algorithm for Clustering Categorical Data . *Journal of Computer Science and Technology*, vol. 17, no. 5, pp. 611-625, 2002.

[2] Jiawei Han, Jian Pei, Micheline Kamber, "Data Mining: Concepts and Techniques", Elsevier, 3rd edition, 2011, 744 p.

[3] Alexandros Nanopoulos, Yannis Theodoridis, Yannis Manolopoulos: "C2P:Clustering Based on Closest Pairs . *Proc. 27th Int'1 Conf. On Very Large Database*, Rome Italy, September 2001.

[4] GuojunGan, Chaoqun Ma, Jianhong Wu, "Data Clustering: Theory, Algorithms, and Applications", ASA-SIAM Series on Statistics and Applied Probability, 2007.

[5] M. Ester, H. P. Kriegel, J. Sander, X. Xu: "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases . *Proc. 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD'96)*, pp. 226-231, Portland OR, Aug. 1996.

[6] M. K Ng, M. J Li, Z. X Huang, Z. Y He "On the impact of dissimilarity measure in k-modes clustering algorithm." IEEE transactions on Pattern Analysis and Machine Intelligence 29 (3) (2007) 503-507.

[7] T. Zhang, R. Ramakishnan, M. Livny: "BIRTH: An Efficient Data Clustering Method for Very Large Databases . *Proc. of the ACM_SIGMOD Int'1 Conf. Management of Data*, 1996, pp. 103-114.

[8] D. W Kim, K. H Lee, D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids", Pattern recognition letters 25 (2004) 1263-1271.

[9] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim: "CURE: A Clustering Algorithm for Large Databases . *Proc. of the ACM SIGMOD Int'1 Conf. Management of Data*, 1998, pp. 73-84.

[10] G. Karypis, E.-H. Han, V. Kumar: "CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling . *IEEE Computer*, Vol. 32, No. 8, 68-75, 1999.

[11] G. Sheikholeslami, S. Chatterjee, A. Zhang: "WaveCluster: A Multi-resolution Clustering Approach for VeryLarge Spatial Databases . *Proc. 1998 Int. Conf. On Very Large Databases*, pp. 428-439, New York, August, 1998.

[12] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim: ROCK:A Robust Clustering Algorithm for Categorical Attributes . *In Proc. 1999 Int. Conf. Data Engineering*, pp. 512-521, Sydney, Australia, Mar.1999.

[13] David Gibson, Jon Kleiberg, Prabhakar Raghavan: "Clustering Categorical Data: An Approach Based on Dynamic Systems . *Proc. 1998 Int. Conf. On Very Large Databases*, pp. 311-323, New York, August 1998.

[14] Yi Zhang, Ada Wai-chee Fu, Chun Hing Cai, Peng-Ann Heng: "Clustering Categorical Data . *In Proc. 2000 IEEE Int. Conf. Data Engineering*, San Deigo, USA, March 2000.