Journal of Computer Science 7 (11): 1639-1645, 2011
ISSN 1549-3636
© 2011 Science Publications

An Integrated Framework for Mixed Data Clustering Using Self Organizing Map

[1]Hari Prasad Devaraj and [2]M. Punithavalli
[1]Sri Ramakrishna Institute of Technology,
[2]Department of Computer Applications,
SNS College of Engineering, Coimbatore, India

Abstract: Problem statement: Clustering plays an important role in data mining of large data and helps in analysis. This develops a vast importance in research field for providing better clustering technique. There are several techniques exists for clustering the similar kind of data. But only very few techniques exists for clustering mixed data items. This leads to the requirement of better clustering technique for classification of mixed data. The cluster must be such that the similarity of items within the clusters is increased and the similarity of items from different clusters must be reduced. The existing techniques possess several advantages and at the same time various disadvantages also exists. **Approach:** To overcome those drawbacks, Self-Organizing Map (SOM) and Extended Attribute-Oriented Induction (EAOI) for clustering mixed data type data can be used. This will take more time for clustering. A modified SOM was proposed based on batch learning. **Results:** The experimentation for the proposed technique was carried with the help of UCI Adult Data Set. The number of clusters resulted for the proposed technique is lesser when compared to the usage of SOM. Also the outliers were not obtained by using the proposed technique. **Conclusion:** The experimental suggests that the proposed technique can be used to cluster the mixed data items with better accuracy of classification.

Key words: Attribute-oriented induction, clustering technique, data mining, training pattern, self-organizing map, batch learning, Better Matching Unit (BMU), numeric attributes, scientific data analysis

INTRODUCTION

One of the widely used techniques in data mining (Wang *et al*., 2010) is clustering. It focuses on grouping a whole data based on its similarity measures that depends on some distance measure. Most techniques of clustering comprise document grouping, scientific data analysis and customer/market segmentation. In knowledge discovery, a basis data mining technique used is data clustering. For investigative data, the clustering with the help of Gaussian mixture models is widely used. The six sequential, iterative steps of Data mining processes are:

- Problem definition
- Data acquisition
- Data preprocessing and survey
- Data modeling
- Evaluation
- Knowledge deployment

The intention of analysis before data preprocessing is to achieve close knowledge into the data possibilities and troubles to find whether the data are enough.

The basic process in data mining (Chen *at el*., 2010) (Syurahbil *et al*., 2009) is the construction of clusters or homogenous category by dividing the set of objects in the databases. It is highly useful in various purposes like classification aggregation and segmentation or dissection.

Generally, clustering includes the classification of the provided data that includes n points in m dimension into k clusters. The clustering (Gandaseca *et al*., 2009) must be such that the data points in the respective cluster (Onoghojobi, 2010) should be highly identical to one another. The troubles involved clustering techniques are: Identifying a likeness measure to guess the similarity among various data, it is hard to determine the appropriate techniques for identifying the identical data in unsupervised way and derive a description that can distinguish the data of a cluster in

Corresponding Author: Hari Prasad Devaraj, Department of Computer Applications, Sri Ramakrishna Institute of Technology, India

an efficient way. Euclidean distance measure is helpful in existing clustering techniques for identifying the similarity among various data. When the data items are categorical or mixed, the similarity cannot be identified by Euclidean distance measure. For the data gathered from banks, or health sector, web-log data and biological sequence data which are categorical data require better clustering technique (Affendey *et al*., 2010) (Nassiry *et al*., 2009). It is highly difficult to cluster the categorical data into meaningful category with good distance measure, capturing adequate data similarity and to utilize conjunction with an efficient clustering algorithm (Hari Prasad and Punithavallim, 2010a).

For dealing with mixed numeric and categorical data, only few techniques exist. One of the techniques is usage of Self-Organizing Map (SOM) and Extended Attribute-Oriented Induction (EAOI) for clustering mixed data type. This will take more time for clustering. To overcome this, a modified SOM is proposed in this study based on batch learning.

Related works: The different existing clustering techniques are discussed here which is proposed by different authors.

Roy and Sharma (2010) proposed a Genetic K-Means Clustering Algorithm for Mixed Numeric and Categorical Data Sets. A novel method was put forth by Juha *et al*. for clustering of Self-Organizing Map. According to the method proposed in this study the clustering is carried out using a two-level approach, where the data set is first clustered using the SOM and then, the SOM is clustered.

Customer behavior pattern is discovered based on Mixed Data Clustering is proposed by Mingzhi *et al*. (2009) To be effective to retain customers and enhance the marketing capabilities, it is necessary to improve the personalization of e-commerce systems. Clustering is a reliable and efficient technology to provide personal service in e-commerce system. However, current research on clustering algorithm usually based on numeric data or categorical data. To analysis customer behavior, mixed data set must be handled. With extending the ROCK algorithm, a novel method to deal with mixed data set was proposed and experiment shows the new algorithm is efficient and successful.

Al-Shaqsi and Wang (2010) put forth a clustering ensemble technique for clustering mixed data. This study provides a clustering ensemble approach based on novel three-staged clustering algorithm. A clustering ensemble is a model that looks for to best merge the outputs of numerous clustering algorithms with a decision fusion function to attain a more accurate and stable final output. The ensemble is built with the proposed clustering approach as a core modeling technique that is used to produce a sequence of clustering outcomes with different conditions for a particular dataset. Then, a decision aggregation system such as voting is utilized to discover a combined partition of the different clusters. The voting mechanism takes only experimental outcomes that generate intra-similarity value better when compared to the average intra-similarity value for a specified interval. The objective of this process is to obtain a clustering outcome that reduces the number of disagreements among different clustering outcome. The ensemble approach has been evaluated on 11 benchmark datasets and evaluated with some individual techniques including TwoStep, k-means, squeezer, k-prototype and some ensemble based methods including k-ANMI, ccdByEnsemble, SIPR and SICM. The experimental results showed its strengths over the compared clustering algorithms.

A Comparative Study in Classification Techniques for Unsupervised Record Linkage Model is proposed by Ektefa *et al*. (2011).

Efficient ensemble approach for mixed numeric and categorical data is recommended by Reddy and Kavitha (2010). The majority of previous clustering algorithms concentrate on numerical data whose inbuilt geometric characteristics can be exploited obviously to define distance functions between data points. On the other hand, a large amount of the data present in the databases is categorical, where attribute values will not be logically ordered as numerical values. Because of the differences in the characteristics of these two categories of data, efforts to build up criteria functions for mixed data have been not very successful. In this paper, proposed a novel divide-and-conquer method to solve this difficulty. Initially, the real mixed dataset is segmented into two sub-datasets: The pure categorical dataset and the pure numeric dataset. Subsequently, accessible well recognized clustering approaches designed for different types of datasets are utilized to generate equivalent clusters. In the end, the clustering outcome on the categorical and numeric dataset are integrated as a categorical dataset, on which the categorical data clustering approach is utilized to obtain the final result. The major involvement in this research is to present an algorithm framework for the mixed attributes clustering difficulty, in which existing clustering algorithms can be effortlessly incorporated.

MATERIALS AND METHODS

Modified self-organizing map: The Self-Organizing Map (SOM) (Zamani *et al*., 2009) is an unsupervised neural network (Abghari *et al*., 2009) (Ali *et al*., 2009) (Qicai *et al*., 2009) that assigns high-dimensional data onto a low dimensional grid, generally two-dimensional and conserves the topological connection of the original data. In other words, similar data inclined to collect mutually on a trained map. Training an SOM usually involves two phase: The identifying and the adjusting steps. In the identifying phase, every training pattern contains the units of the map and finds the Better Matching Unit (BMU) that is highly identical to the training model. Next, in the adjusting phase, the BMU and its neighbors are updated to be similar to the training pattern. Repeat these two phases for all patterns in the training data set until the map converges.

The identifying and the adjusting phases can be described by the following formulas:

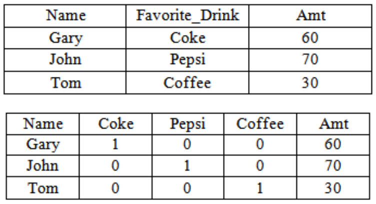
$$\|x - m_b\| = \min_c\{\|x - m_c\|\}$$

$$m_c(t-1) = m_c(t) + a(t)h_{bc}(t)[x(t) - m_c(t)]$$

where, x represents the input vector. m_c and m_b represents the model vectors of unit c and BMU, correspondingly. $\alpha(t)$ and $h_{bc}(t)$ represents the learning rate and the neighborhood function, both decreasing steadily upon increasing the training step t.

As it was initially provided, SOM has several applications. In its beginning stages, the application is usually based on engineering. Later in current trend, usage to data mining and other fields have arrived.

Due to the ability of topology protection, SOM is an outstanding technique in the exploratory stage of data mining and has currently been combined with existing clustering techniques to assist in cluster analysis. It has been determined that the combined technique decrease computation time and carry out better in comparison with other direct clustering techniques.



Name	Favorite_Drink	Amt
Gary	Coke	60
John	Pepsi	70
Tom	Coffee	30

Name	Coke	Pepsi	Coffee	Amt
Gary	1	0	0	60
John	0	1	0	70
Tom	0	0	1	30

Fig. 1: The conventional approach transforms the categorical attribute Favorite_Drink to three binary attributes with domain {0, 1} prior to training an SOM

One of the demerits of SOM is overfitting. The conventional SOM cannot straightly handle categorical attributes. Finding the BMU of a training pattern usually resorts to computing the Euclidean distance, thus, only appropriate for numeric data. For mixed data, binary transformation that changes each mixed data to a set of binary attributes (e.g., Fig. 1) is usually carried out before the training phase.

In machine-learning, the working of a trained technique is usually expressed in its generalization performance, i.e., its capacity to perform perfectly new data not included in the training set. When the performance of the trained model is much lesser than its performance on the training material, overfitting is considered. Overfitting is resulted because of the sparseness of the training material. The next reason cause for overfitting may be a high degree of non-linearity in the training material. In those situations, the learning technique may not be able to learn more from the training data than the classification of the training instances itself.

On the other hand, the binary transformation technique has at least four demerits:

- Similarity details between categorical values are not conveyed
- When the domain of a categorical attribute is higher, the transformation maximizes the dimensionality of the transformed relation, resulting in wasting storage space and in maximizing the training time
- Maintenance is very complex; when the attribute domain is modified, the new relation scheme requires modifying also
- The names of binary attributes unsuccessful to conserve the semantics of the original categorical attribute

Another common technique for handling categorical values in clustering technique is simple matching, where a comparison of two identical values results in a difference 0; otherwise, two distinct values result in a difference 1. On the other hand, this technique does not consider the similarity between categorical values into consideration and, hence, may fail to faithfully disclose the structure of mixed data.

In order to overcome these issues, batch learning is used in this study.

Batch-learning for self-organizing map:

Initializing of the weight vectors: The initial weight vectors $w_{ij}^{(init)}$ are determined according to the principal component examination of the input vectors x_k as below:

$$w_{ij}^{(init)} = \overline{x} + 5\sigma_1 \frac{i-1/2}{1} b_1 + 5\sigma_2 \frac{j-j/2}{j} b_2$$

Here, $\overline{x}$ represents the average vector of x_k, b_1 and b_2 are eigen vectors for the first and second principal components and σ_1 and σ_2 are the standard deviations of the first and second principal components. The second dimension J is defined by $J = [\sigma_2/\sigma_1 I]$.

Classifying of the input vectors: The distances between the input vector x_k and the weight vector w_{ij} are computed and x_k is categorized into the weight vector $Wi'j'$ with the least distance.

Updating of the weight vectors: The $_{ij}$th weight vector is updated with:

$$w_{ij}^{(init)} = w_{ij}^{(old)} + a(t)\left\{\frac{\sum_{xkesij} x_k}{N_{ij}} - w_{ij}^{(old)}\right\}$$

Here, the components of set S_{ij} are input vectors classified into $Wi'j'$ satisfying $i\text{-}\beta(t) \le j` \le j\beta(t)$ and $j\text{-}\beta(t) \le j` \le j+\beta(t)$ and N_{ij} are the numbers of components of S_{ij}. The two parameters $a(t)(0<a(t)<1)$ and $\beta(t)(0\le\beta(t))$ are learning coefficients for the tth cycle defined by:

$$a(t)\max = \left\{0.001, a_{init}(1-\frac{t}{\tau_a})\right\}$$

$$\beta(t)\max = \left\lfloor \left\{0.001, \beta_{init}(1-\frac{t}{\tau_\beta})\right\} \right\rfloor$$

where, α_{init} and β_{init} are the initial values and τ_α and τ_β are the time constants.

Extended attribute-oriented induction: To trounce the drawback of major values and numeric attributes, an extension to the conventional AOI is proposed in this study. This provides the ability of exploring the major values and a choice for processing numeric attributes. For the exploration of major values, a parameter majority threshold β is introduced. If certain values (i.e., major values) take up a chief portion (exceeding β) of an attribute, the extended AOI (EAOI) conserves those chief values and generalizes other non major values. If no major values present in an attribute, the EAOI proceeds like the AOI, generating the same results as that of the conventional approach. Furthermore, if β is set to 1, the EAOI degenerates to the AOI.

For solving the problems of constructing subjectively numeric concept hierarchies and generalizing boundary values, an alternative for processing numeric attributes is proposed: Users can choose to compute the average and deviation of the aggregated numeric values instead of generalizing those values to discrete concepts. Under this alternative, only categorical attributes are generalized. The average and deviation of numeric attributes of the merged tuples are calculated and then replace the original numeric values. The computed deviation reveals the dispersion of numeric values; the less the deviation is, the more concentrated the values are; otherwise, the more diversified the values are.

The EAOI algorithm is outlined as follows:

Algorithm: An extended attribute-oriented induction algorithm for major values and alternative processing of numeric attributes

Input: A relation W with an attribute set A; a set of concept hierarchies; generalization threshold and majority threshold.

Output: A generalized relation P.

Method:

- Determine whether to generalize numeric attributes
- For each attribute Ai to be generalized in W
- Determine whether Ai should be removed and if not, determine its minimum desired generalization level Li in its concept hierarchy
- Construct its major-value set M_i according to and
- For construct the mapping pair as otherwise, as (v,v)
- Derive the generalized relation P by replacing each value v by its mapping value and computing other aggregate values

In Step 1, if numeric attributes are not to be generalized, their averages and deviations will be calculates in Step 3. Step 2 intends at preparing the mapping pairs of attribute values for generalization. First, in Step 2.1, an attribute is eliminated either because there is no concept hierarchy defined for the attribute, or its higher-level concepts are expressed in terms of other attributes. In Step 2.2, the attribute's major-value set Mi is constructed, which consists of the first a(<θ) count leading values if they take up a major portion (≥β) of the attribute, where θ is the generalization threshold that sets the maximum number of distinct values allowed in the generalized attribute.

In Step 2.3, if v is one of the major values, its mapping value remains the same, i.e., major values will not be generalized to higher-level concepts. Otherwise, v will be generalized by the concept at level Li by excluding the values contained in both the major-value set and the leaf set of the v_{Li} subtree (i.e., v_{Li}-M_{Li} where M_{Li} = Leaf(v_{Li})$\cap$ M_i). Note that, if there are no major values in A_i, M_i and M_{Li} will be empty. Accordingly, the EAOI will behave like the AOI. In Step 3, aggregate values are computed, including the accumulated count of merged tuples, which have identical values after the generalization and the averages and deviations of numeric attributes of merged tuples if numeric attributes are determined not to be generalized.

RESULTS

The proposed clustering technique is experimented with UCI Adult Data Set. The data set contains 15 attributes that include eight categorical, six numerical and one class attributes. 10, 000 tuples from the 48,842 tuples are chosen randomly for the evaluation.

For the attribute choosing, the method of relevance analysis based on information gain is utilized. The relevance threshold was set to 0.1 and seven qualified attributes are obtained: Marital-status, Relationship, Education, Capital_gain, Capital_loss, Age and Hours_per_week. The first three are categorical and the others are numeric.

The map volume is 400 units. The training parameters are set to the same with that of the earlier experimentation.

The number of resultant clusters by using SOM and modified SOM with different distance criteria is provided in Table 1 and Fig. 2. It can be seen that the proposed technique results in better categorization.

To evaluate how the clustering improves the likelihood of similar values falling in the identical cluster, the average categorical utility of clusters can be helpful. The categorical utility function aims to increase both the probability that the two data in the same cluster have attribute values in common and the probability that the data from various clusters have different values. The higher the value of categorical utility, the better the clustering fares. The average categorical utility of a set of clusters is calculated as follows:

$$\mathrm{ACU} = \frac{1}{k}\sum_{k}\left(\frac{|C_k|}{|D|}\sum_{i}\sum_{j}[P(A_i = V_i \,|C_k)^2 - P(A_i = V_{ij})]^2\right)$$

where $P(A_i$ = Vij$|C_k)$ is the conditional probability that the attribute A_i has the values V_{ij} given the cluster C_k and $P(A_i = V_{ij})$ is the overall probability of Ai having V_{ij} in the entire data set.

Table 1: Number of resultant clusters for using SOM and modified SOM with different distance criteria

	Modified-SOM		SOM	
	Cluster	Outliers	Cluster	Outlier
D = 0	81	-	88	-
d≤1.414	12	-	19	-
d≤2.828	5	-	9	-
d≤2 and A_{dj}	6	5	14	1

Table 2: The increased rate of the average CU and the expected entropy of the class attribute salary

	Modified SOM			
	Leaf level	Level1	Increased	Exp.entropy
d≤1.414	0.099	0.102	17%	0.601
d≤2.828	0.158	0.189	19%	0.678
d≤3 and Adj	0.171	0.197	22%	0.602
SOM				
d≤1.414	0.118	0.108	-.07%	0.613
d≤2.828	0.119	0.107	-0.08%	0.701
d≤3 and Adj	0.121	0.106	-0.09%	0.613

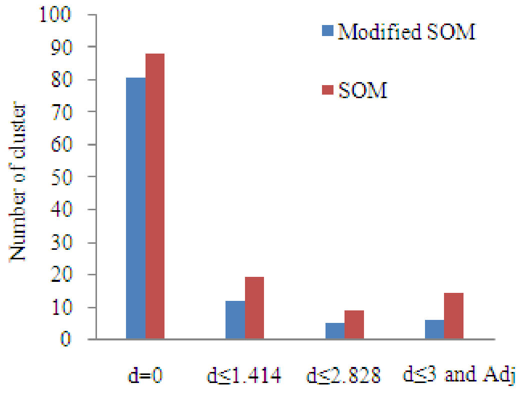


Fig. 2: Number of resultant clusters for using SOM and modified SOM with different distance criteria

The ACU of categorical values of clusters formed by the three clustering criteria are computed at the leaf level and Level 1 of the distance hierarchies and the improved rate, as shown in Table 2. The ACU at Level 1 is computed by generalizing categorical values to their values at Level 1 and then applying distance function.

The larger increased rates in the BP-SOM approach indicate that the BP-SOM influences the clustering in the way of helping group similar categorical values together, where the similarity is defined via distance hierarchies. In addition, compared to those of the SOM, the spots of the BP-SOM spread less widely. This also indicates the effect of taking the similarity between categorical values into consideration during training.

DISCUSSION

There are several problems exists in the existing clustering algorithms especially while clustering the data with mixed data types. This study analyzes those existing techniques and comes with the new technique for clustering the mixed data items. The Modified Self Organizing Map is used in this study for better classification of data. This Modified SOM uses batch learning procedure in its leaning algorithm. Then for constructing the hierarchy of generalized relations, Extended Attribute-Oriented Induction is used in this study. The combination of these two results in better clustering of the mixed data. The experiment result indicates that the proposed technique clusters the mixed data effectively.

CONCLUSION

This study focuses on efficient clustering technique for mixed category data. There are different technique exist for clustering categorical, but all those technique resulted in several disadvantages. To overcome this issue, the clustering in mixed data can be performed based on Self-Organizing Map and Extended Attribute-Oriented Induction (EAOI). But this technique also takes more time for classification. To this issue, a new modified SOM technique is used in this study based on batch learning. The experiment is performed with the help of UCI Adult Data Set and it can be observed that the better classification result is obtained for the proposed technique when compared to the existing techniques.

REFERENCES

Abghari, H., M. Mahdavi, A. Fakherifard and A. Salajegheh, 2009. Cluster analysis of rainfall-runoff training patterns to flow modeling using hybrid RBF networks. Asian J. Applied Sci., 2: 150-159. DOI: 10.3923/ajaps.2009.150.159

Affendey, L.S., I.H.M. Paris, N. Mustapha, M.N. Sulaiman and Z. Muda, 2010. Ranking of influencing factors in predicting students academic performance. Inf. Technol. J., 9: 832-837. DOI: 10.3923/itj.2010.832.837

Ali, W.H., A.N. Abdalla and W.H. Ali, 2009. Winner-take-all neural network with massively optoelectronic interconnections. Am. J. Applied Sci., 6: 268-272. DOI: 10.3844/ajassp.2009.268.272

Al-Shaqsi, J. and W. Wang, 2010. A clustering ensemble method for clustering mixed data. The 2010 Proceedings of the International Joint Conference on Neural Networks, July 18-23, IEEE Xplore Press, Barcelona, pp: 1-8. DOI:10.1109/IJCNN.2010.5596684

Chen, T.S., J. Chen and Y.H. Kao, 2010. A novel hybrid protection technique of privacy-preserving data mining and anti-data mining. Inform. Technol. J., 9: 500-505. DOI: 10.3923/itj.2010.500.505

Ektefa, M., F. Sidi, H. Ibrahim, M.A. Jabar and S. Memar, 2011. A comparative study in classification techniques for unsupervised record linkage model. J. Comput. Sci., 7: 341-347. DOI: 10.3844/jcssp.2011.341.347

Gandaseca, S., J. Sabang, O.H. Ahmed and N.M.A. Majid, 2009. Vegetation assessment of peat swamp forest using remote sensing. Am. J. Agric. Biol. Sci., 4: 167-172. DOI: 10.3844/ajabssp.2009.167.172

Hari Prasad, D and M. Punithavallim, 2010a. A review on data clustering algorithms for mixed data. Global J. Comput. Sci. Technol., 10: 43-48.

Mingzhi, C., X. Yang, T. Yangge, W. Cong and Y. Yixian, 2009. Customer behavior pattern discovering based on mixed data clustering. Proceedings of the International Conference on Computational Intelligence and Software Engineering, IEEE Xplore Press, Wuhan, pp: 1-4. DOI: 10.1109/CISE.2009.5366556

Nassiry, M.R., A. Javanmard and R. Tohidi, 2009. Application of statistical procedures for analysis of genetic diversity in domestic animal populations. Am. J. Anim. Vet. Sci., pp: 136-141. DOI: 10.3844/ajavsp.2009.136.141

Onoghojobi, B., 2010. Subsample goal model for multihalver on outliers. J. Math. Stat., 6: 347-349. DOI: 10.3844/jmssp.2010.347.349

Qicai, L., Z. Kai, Z. Zehao, F. Lengxi and O. Qishui et al., 2009. The use of artificial neural networks in analysis cationic trypsinogen gene and hepatitis b surface antigen. Am. J. Immunol., 5: 50-55. DOI: 10.3844/ajisp.2009.50.55

Reddy, M.V.J. and B. Kavitha, 2010. Efficient ensemble algorithm for mixed numeric and categorical data. Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research, Dec. 28-29, IEEE Xplore Press, Coimbatore, pp: 1-4. DOI: 10.1109/ICCIC.2010.5705738

Roy, D.K. and L.K. Sharma, 2010, Genetic K-means clustering algorithm for mixed numeric and categorical data sets. International Journal of Artificial Intelligence and Applications, 1: 23-28.

Syurahbil, N.A., M.F. Zolkipli and A.N. Abdalla, 2009. Intrusion preventing system using intrusion detection system decision tree data mining. Am. J. Eng. Applied Sci., 2: 721-725. DOI: 10.3844/ajeassp.2009.721.725

Wang, Y.H., D.A. Chiang, S.W. Lai and C.J. Lin, 2010. Applying data mining techniques to WIFLY in customer relationship management. Inform. Technol. J., 9: 488-493. DOI: 10.3923/itj.2010.488.493

Zamani, A., M. Nedaei and R. Boostani, 2009. Tectonic zoning of Iran based on self-organizing map. J. Applied Sci., 9: 4099-4114. DOI: 10.3923/jas.2009.4099.4114