ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Clustering of Data with Mixed Attributes based on Unified Similarity Metric

M.Soundaryadevi[1], Dr.L.S.Jayashree[2]

Dept of CSE, RVS College of Engineering and Technology, Coimbatore, Tamilnadu, India[1]

Professor & Dean, Dept of CSE, RVS College of Engineering and Technology, Coimbatore, Tamilnadu, India[2]

ABSTRACT— Most of the clustering approaches are applicable to purely numerical data or purely categorical data but not both. There exists an awkward gap between the similarity metrics for categorical and numerical data, so it is a non trivial task for clustering of data with mixed attributes. A general clustering algorithm for based on object cluster similarity is framed which clusters the data with mixed attributes. Moreover, clustering techniques are applied in Educational Data Mining (EDM) to group of students according to their customized features. Student data set is a collection of students' personal characteristics, skill profiles, etc. It mostly contains the collection of attributes of different types. So, computationally efficient and simple clustering algorithm is required for this kind of data sets. The group of students obtained from clustering technique can be used by the instructor to build an effective learning system, to promote group learning, to provide adaptive contents etc. Here an iterative clustering algorithm based on object cluster similarity (OCIL) is to be applied on the student data set with mixed attributes. This will serve as a valid guide line for development of intelligent tutoring system.

KEYWORDS- Clustering; Categorical; Numerical

I. INTRODUCTION

With the amazing progress of both computer hardware and software, a vast amount of data is generated and collected daily. There is no doubt that data are meaningful only when one can extract the hidden information inside them. However, "the major barrier for obtaining high quality knowledge from data is due to the limitations of the data itself". These major barriers of collected data come from their growing size and versatile domains. Thus, data mining that is to discover interesting patterns from large amounts of data within limited sources (i.e., computer memory and execution time) has become popular in recent years.

Clustering is considered an important tool for data mining. The goal of data clustering is aimed at dividing the data set into several groups such that objects have a high degree of similarity to each other in the same group and have a high degree of dissimilarity to the ones in different groups. Each formed group is called a cluster. Useful patterns may be extracted by analyzing each cluster. For example, grouping customers with similar characteristics based on their purchasing behaviors in transaction data may find their previously unknown patterns. The extracted information is helpful for decision making in marketing.

Various clustering applications have emerged in diverse domains. However, most of the traditional clustering algorithms are designed to focus either on numeric data or on categorical data. The collected data in real world often contain both numeric and categorical attributes. It is difficult for applying traditional clustering algorithm directly into these kinds of data. Typically, when people need to apply traditional distance-based clustering algorithms to group these types of data, a numeric value will be assigned to each category in this

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

attributes. Some categorical values, for example "low", "medium" and "high", can easily be transferred into numeric values. But if categorical attributes contain the values like "red", "white" and "blue", etc., it cannot be ordered naturally. Under the circumstances, a straightforward way to overcome this problem is to transform the categorical values into numerical ones, e.g. the binary strings, and then apply the aforementioned numerical-value based clustering methods. Nevertheless, such a method has ignored the similarity information embedded in the categorical values and cannot faithfully reveal the similarity structure of the data sets. Hence, it is desirable to solve this problem by finding a unified similarity metric for categorical and numerical attributes such that the metric gap between numerical and categorical data can be eliminated. Subsequently, a general clustering algorithm which is applicable to numerical and categorical data can be presented based on this unified metric. During the past decades, some works which try to find a unified similarity metric for categorical and numerical attributes. However, a computationally efficient similarity measure remains to be developed.

II. LITERATURE SURVEY

Several methods have been presented which can be grouped into two lines. In the first line, the algorithms are essentially designed for purely categorical data, although they have been applied to the mixed data as well by transforming the numerical attributes to categorical ones via a discretization method. Along this line, several methods have been proposed based on the perspective of similarity metric, graph partitioning or information entropy.
Li and Biswas [5] presented the Similarity Based Agglomerative Clustering (SBAC) algorithm which is based on Goodall similarity metric that assigns a greater weight to uncommon feature value matching in similarity computations without the prior knowledge of the underlying distributions of the feature values.

Similarity-Based Agglomerative Clustering (SBAC) algorithm that works well for data with mixed numeric and nominal features. An agglomerative algorithm is employed to construct a dendrogram and a simple distinctness heuristic is used to extract a partition of the data. The performance of SBAC has been studied on real and artificially generated data sets. Results demonstrate the effectiveness of this algorithm in unsupervised discovery tasks. Comparisons with other clustering schemes illustrate the superior performance of this approach. This method has a good capability of dealing with the mixed attributes, but its computation is quite laborious.

Bindiya M Varghese , Jose Tomy J,[7]" Clustering Student Data to Characterize Performance Patterns", Over the years the academic records of thousands of students have accumulated in educational institutions and most of these data are available in digital format. Mining these huge volumes of data may gain a deeper insight and can throw some light on planning pedagogical approaches and strategies in the future. They proposed to formulate this problem as a data mining task and use k-means clustering and fuzzy c-means clustering algorithms to evolve hidden patterns and compared the performance of K-Means and Fuzzy C Means to cluster the student data to analyze the performance based on their skills.

Oyelade O. J, Oladipupo O. O [6]," Application of k-Means Clustering algorithm for prediction of Students' Academic Performance" implemented K-Means algorithm for prediction of students academic performance based on the GPA. They clustered the data set according to their performance levels such as Poor, Fair, Average, Good, Excellent. They also analyzed the performance as the cluster size increases. This method takes into account only a single factor and it is also a numerical attribute.

Ming-Yi ShihA,[3] "Two-Step Method for Clustering Mixed Categorical and Numeric Data", In this approach the items in categorical attributes are processed to construct the similarity or relationships among them based on the ideas of co-occurrence; then all categorical attributes can be converted into numeric attributes based on these constructed relationships. Finally, since all categorical data are converted into numeric, the existing clustering algorithms can be applied to the dataset without pain. Nevertheless, the existing clustering algorithms suffer from some disadvantages or weakness,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

the proposed two-step method integrates hierarchical and partitioning clustering algorithm with adding attributes to cluster objects. This method defines the relationships among items, and improves the weaknesses of applying single clustering algorithm. The TMCM algorithm integrates HAC and k-means clustering algorithms to cluster mixed type of data. Applying other algorithms or sophisticated similarity measures into TMCM may yield better results.

III. UNIFIED SIMILARITY METRIC FOR CLUSTERING MIXED DATA

Yiu-ming Cheung, Hong Jia,[1] proposed the following metric for calculation of similarity. The general task of clustering is to classify the given objects into several clusters such that the similarities between objects in the same group are high while the similarities between objects in different groups are low. For categorical attribute, each attribute represents an important feature of the given object. Clustering or classification analysis can be conducted on categorical attributes similar to decision tree method. So the categorical attributes are treated individually and the numerical part is treated as a whole. The number of features that contributes to clustering analysis will be d_c+1. (i.e d_c categorical features and 1 numerical vector).Let the object cluster similarity between x_i and cluster C_j denoted as $S(x_i, C_j)$,be the average of similarity calculated based on each feature, we will have

$$S(x_i, C_j) = \frac{1}{d_f}\left[S(x_{i1}^c, C_j) + S(x_{i2}^c, C_j) + \cdots S(x_{id_c}^c, C_j)\right] + \frac{1}{d_f} S(x_i^u, C_j)$$

In practice due to different distributions of each categorical attributes each often have unequal importance for clustering analysis. Therefore the equation for similarity of categorical attribute can be modified with

$$S(x_i^c, C_j) = \sum_{r=1}^{d_c} w_r\, S(x_{ir}^c, C_j)$$

where $w_r$is the weight of categorical attribute A_r satisfying $0 \leq w_r \leq 1$ and $\sum_{r=1}^{d_c} w_r = 1$.
The similarity between a categorical attribute value x_{ir}^c and cluster C_j , i={1,2,…,N}, r={1,2,…, d_c}and j={1,2,…,k} is defined as

$$S(x_{ir}^c, C_j) = \frac{\sigma_{A_r = x_{ir}^c}(C_j)}{\sigma_{A_r \neq \mathrm{NULL}}(C_j)}$$

where $\sigma_{A_r = x_{ir}^c}(C_j)$ counts the number of objects in C_j that have value x_{ir}^c for attribute A_r and $\sigma_{A_r \neq \mathrm{NULL}}(C_j)$ means the number of objects in cluster C_j that have attribute A_r whose value is not equal to NULL.
Suppose the mixed data x_i with d different attributes and consists of d. The object-cluster similarity metric for mixed data is defined as

$$S(x_i, C_j) = \frac{d_c}{d_f} S(x_i^c, C_j) + \frac{1}{d_f} S(x_i^u, C_j)$$

$$S(x_i, C_j) = \frac{d_c}{d_f} \sum_{r=1}^{d_c} \left(\frac{H_{A_r}}{\sum_{t=1}^{d_c} H_{A_t}} \cdot \frac{\sigma_{A_r = x_{ir}^c}(C_j)}{\sigma_{A_r \neq \mathrm{NULL}}(C_j)} \right) + \frac{1}{d_f} \frac{\exp\left(-0.5 \mathrm{Dis}(x_i^u, C_j)\right)}{\sum_{t=1}^{k} \exp\left(-0.5 \mathrm{Dis}(x_i^u, C_t)\right)}$$

where i = 1,2, . . . ,N, j = 1,2, . . . ,k. It can be seen that the defined similarities for categorical and numerical attributes in Eq. (4) are in the same scale. That is, the values for $S(x_i^c, C_j)$ and $S(x_i^u, C_j)$ are within the interval [0,1]. Hence, additional parameters to control the proportions of numerical and categorical distances are not needed any more

IV. RESULTS & DISCUSSIONS

The student data set is taken as a sample for testing the performance of clustering algorithm on data set with mixed attributes. The student record contains data of around 670 students. It consists of the following attributes Students' Gender,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Degree, Department in which the student belongs, students' year of completion, Marks obtained by the student in 10th, 12th or Diploma and the CGPA of the student in degree. The attributes standing arrears and history of arrears represents the number of standing arrears of student and a Boolean attribute for storing history of arrears.

The following table shows the cluster centers based on which the data is grouped into 3 clusters Cluster 0, Cluster 1 and Cluster 2.

Students in Cluster 0 are categorized as students with average skills because the degree CGPA, 12th and 10th marks are in between Cluster 1 and cluster2. Cluster 1 is categorized as good. Since their percentage of marks are high compared to other clusters. Most of the students who belong to this category don't have any history of arrears.

Students in Cluster 2 are categorized as bad. They have the lowest percentage of marks and more number of arrears.

Cluster Centers

Attributes	Cluster 0	Cluster 1	Cluster 2
Gender	M	F	M
Degree	B.E	B.E	B.E
Discipline	ECE	ECE	EEE
Year of Completion	2013	2013	2014
10th Standard %	76.7083	84.4322	77.8437
12th Standard or Diploma %	76.827	81.4359	75.7293
Degree CGPA	7.3755	7.9567	6.5029
Standing Arrears	0.8288	0.1078	5.0546
History of Arrears	YES	NO	YES

The following figure shows the graphical representation of the cluster assignments.The interpretation of clustered output is easier when the output is in the form of graph.More students in cluster 2 has history of arrears than those in cluster 0 and cluster 1. Most of students in cluster 1 don't have history of arrears..

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

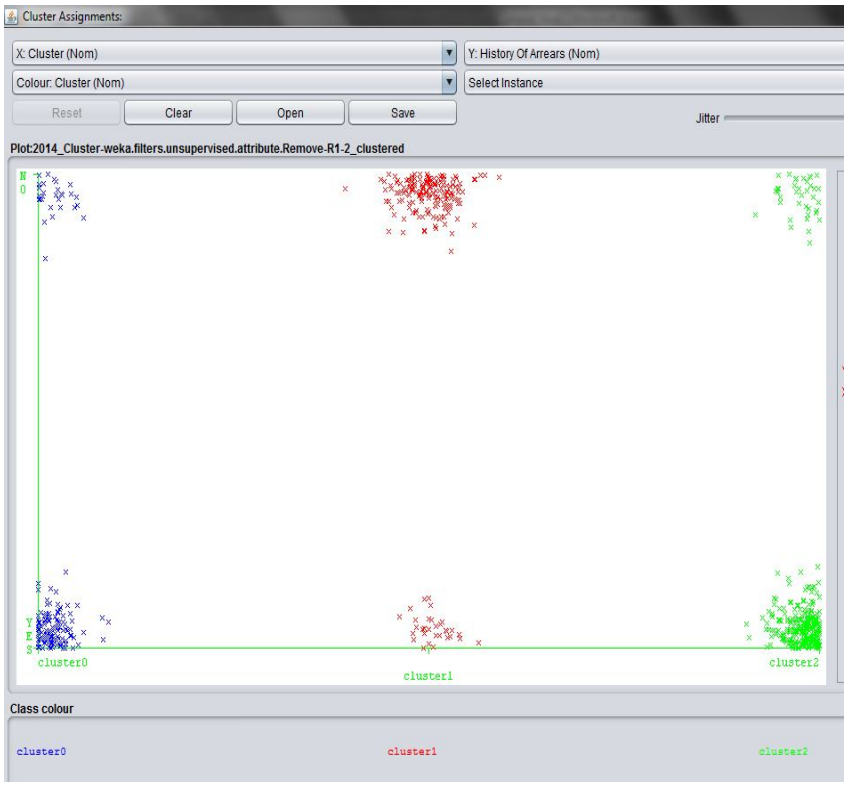

Figure 1 Attribute value distribution among each cluster

The following Figure shows the distribution of cluster values based on the standing arrears and percentage of marks. As the output shows, the students are clustered based on three different skill profiles good, average and below average

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

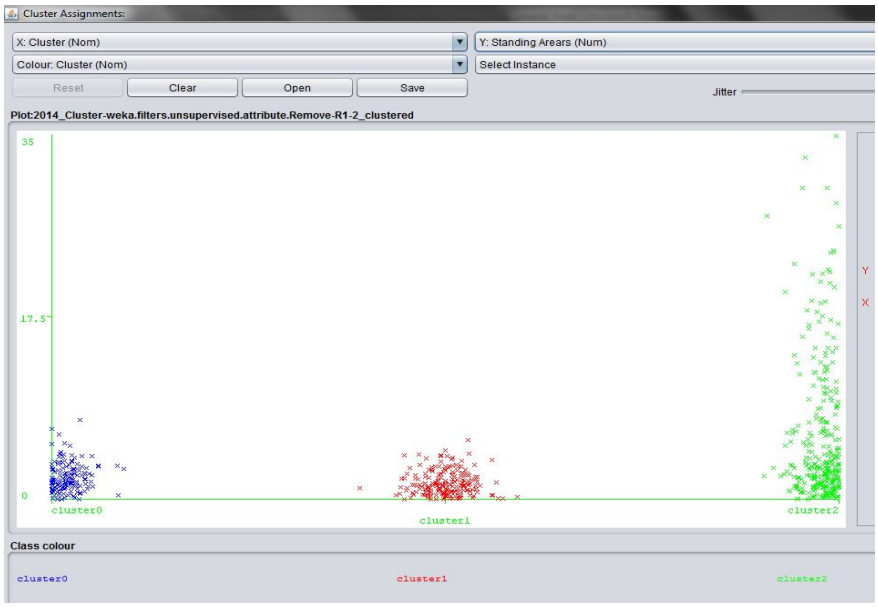

Figure 2 Distribution of Standing arrears among each cluster

V. CONCLUSION

A general clustering framework based on object-cluster similarity, through which a unified similarity metric for both categorical and numerical attributes has been proposed. It is beneficial for clustering on various data types using the above iterative clustering algorithm based on object cluster similarity. The iterative clustering algorithm using object cluster similarity metric effectively clusters the data with both numerical and nominal attributes. As student data is taken as a sample which consists of mixed attributes the students with similar skills are effectively grouped. Accuracy of any clustering algorithm depends on the initialization of cluster centers. As a future work an efficient algorithm for initialization of cluster centers for the algorithm which can cluster mixed numerical and nominal data is to be proposed.

REFERENCES

[1] Yiu-ming Cheung, Hong Jia, "Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number", Elsevier, vol 46,no. 8,pp. 2228–2238, 2013
[2] Cristobal.R and Sebastián.V, "Educational Data Mining: A Review of the State-of-the-Art", IEEE Transactions On Systems, Man, And Cybernetics,2010
[3] David, G. and AmirAverbuch," SpectralCAT: Categorical spectral clustering of numerical and nominal data", Elsevier, vol 45, pp. 416–433, 2012
[4] 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, No. 1, pp. 11_19 ,2010
[5] Han J. and Kamber, M., Data mining : concepts and techniques", Morgan Kaufmann,2012
[6] C.C. Hsu, "Generalizing self-organizing map for categorical data", IEEE Transactions on Neural Networks vol. 17 no. 2, 294–304,2006.
[7] C. Li, G. Biswas, "Unsupervised learning with mixed numeric and nominal data, IEEE Transactions on Knowledge and Data Engineering " vol .14 no.4 673–690, 2002.
[8] Oyelade O. J, Oladipupo O. O," Application of k-Means Clustering algorithm for prediction of Students' Academic Performance" (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence, ,2010.
[9] Varghese, B.M. and Jose Tomy, J., " Clustering Student Data to Characterize Performance Patterns" (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence
[10] SudiptoGuha, Rajeev Rastogi,"ROCK: A Robust Clustering Algorithm for Categorical Attributes"