Categorical Data Clustering using Cosine based similarity for Enhancing the Accuracy of Squeezer Algorithm

R.Ranjani
PG Scholar
Department of Computer Science and Engineering
Sona college of Technology

S.Anitha Elavarasi
Asst Professor
Department of Computer Science and Engineering,
Sona College of Technology

J.Akilandeswari
Professor and Head
Department of Information Technology, Sona College of Technology, Salem 05

ABSTRACT
Clustering categorical data is the major challenge in data mining. Direct comparison of categorical data is not possible as in numerical data, understanding the province of categorical data help to form good quality clusters. This paper presents an algorithm Enhanced Squeezer, which incorporates Data Intensive Similarity Measure for Categorical data (DISC) in Squeezer algorithm. DISC measure, cluster data by understanding domain of the dataset, thus clusters formed are not purely based on frequency distribution as many similarity measures do. Clusters formed using Enhanced Squeezer algorithms are intensive and accurate.

Keywords
DISC Measure, Squeezer, Categorical Data Clustering, Cosine similarity.

1. INTRODUCTION
Clustering is the process of grouping a set of physical objects into classes of similar object or homogenous pattern[12]. In other words clustering is to partition a set of objects into clusters so that objects in the same cluster are more similar between them than objects from other cluster according to the established criterion.

Research community have proposed different clustering algorithm and many are suitable for clustering numerical data. In real world scenario, data in database are categorical in nature, which are raw or unsummarized data, where the attributes cannot be prearranged as numerical values.

Our focus is mainly on clustering categorical data. The concept of similarity or distance for categorical data is not as simple as for continuous data. The typical feature of categorical data is that, the different values that a categorical attribute have cannot be inherently ordered. Therefore, direct comparison of two different categorical values is highly impossible. The concept of similarity for any dataset differs depending on the domain of the dataset.

The idea is to cluster categorical data using Enhanced Squeezer algorithm. The algorithm uses DISC (Data Intensive Similarity Measure for Categorical Data) for similarity computation. DISC identifies the semantics of the data by capturing the relationships that are inherent in the data. Advantages of DISC are it is generic and simple to implement [2].

The paper is organized as follows: Section 1 presents introduction Section 2 presents a Overview of algorithm and measures and the explanation of Squeezer algorithm. Section 3 describes detailed working of proposed system Section 4 describes experimental study. Finally section 5 provides the conclusion of the paper.

2. OVERVIEW OF ALGORITHM AND MEASURES
Existing clustering algorithms can be classified into two main categories: Hierarchical and Partitioning algorithm.

David Arthur et al [6] proposed K-Means++ which is an extension of K-Means algorithm. K-Means++ find the center using probability measure which gives an optimal seed value for the existing K-Means algorithm. The author shows that K-Means++ outperforms K-Means in both speed and accuracy.

Zengyou Ye et al[10] introduced K-ANMI works similar way as K-means algorithm, takes ‘k’ as input and changes class label iteratively for each object to improve the objective function value. Cluster is evaluated in each step using mutual function based criterion-ANMI.

Zengyou Ye et al[11] proposed NabSqueezer algorithm, an improved Squeezer algorithm. NabSqueezer algorithm gives more weight to uncommon attribute value matches for finding similarity in similarity computation of Squeezer algorithm. In this algorithm weight of each attribute is precalculated using More Similar Attribute Value Set (MSFVS) method.

Taoying et al [4] proposed Fuzzy Clustering Ensemble Algorithm for Partitioning Categorical Data makes use of relationship degree of attributes for pruning a part of attributes. Descartes subset is used for finding the cluster membership. Both relationship degree and Descartes subsets are used for establishing the relationship between objet as well as minimizes the objective function.

Z.Huang and M.K. Ng [7] presented Fuzzy K-Modes algorithm makes use of a simple matching dissimilarity measure (Generalized Hamming distance) and Mode values

for clustering the categorical objects. The algorithm uses

to minimize cost function $F_c(X,Z)$ and update z at each iteration.

Wang Jiacai and Gu Ruijun [8] developed Extended Fuzzy K-Means algorithm uses expanded form of cluster centroid vector representation to keep the clustering information and update the method in the same way as in fuzzy k-means.

Aditya Desai [2] use similarity which are neighborhood-based or incorporate the similarity computation into the learning algorithm. These measures compute the neighborhood of a data point but not suitable for calculating similarity between a pair of data instances X and Y.

update method inorder

Rishi Sayal et al [1] proposed a concept called Context Based Similarity Measure which is achieved in relational database through Functional Dependency. The Context Based similarity finds the similarity between components by checking the contexts in which they appear.

2.1 Similarity Measure

Shyam Boriah et al [3] the author presents a comparative study on number of similarity measure, such as Goodall , Occurrence Frequency, Overlap, Inverse Occurrence Frequency, Burnaby, Gambaryan. Few measures are tabulated as follows.

Table 1. List of similarity measure

Measure	Formula	Comments
Gambaryan	$\begin{cases} -[\widehat{p_k}(x_k)\log\widehat{p_k}(x_k) + (1-\widehat{p_k}(x_k))\log_2(1-\widehat{p_k}(x_k))] & if\ x_k = y_k \\ 0 & otherwise \end{cases}$	More weight to matching objects that occurs in between being frequent and rare.
Overlap	$\begin{cases} 1 & if\ x_k = y_k \\ 0 & otherwise \end{cases}$	Counts the number of attributes that match in the two data instances. The disadvantage of overlap measure is that does not distinguish between values taken by the attribute
Occurrence Frequency	$\begin{cases} 1 & if\ x_k = y_k \\ \dfrac{1}{1+\log\frac{N}{f_k(x_k)} * \log\frac{N}{f_k(y_k)}} & otherwise \end{cases}$	Mismatches on less frequent values are assigned lesser similarity and mismatches on more frequent values are assigned higher similarity.
Burnaby	$\begin{cases} 1 & if\ x_k = y_k \\ \dfrac{\sum_{q\in A_k} 2\log(1-\widehat{p_k}(q))}{\log\frac{\widehat{p_k}(x_k)\widehat{p_k}(y_k)}{(1-\widehat{p_k}(x_k))(1-\widehat{p_k}(y_k)) + \sum_{q\in A_k} 2\log(1-\widehat{p_k}(q))}} & otherwise \end{cases}$	The author assigns low similarity to mismatches on rare values and high similarity to mismatches on frequent values.
Inverse Occurrence Frequency	$\begin{cases} 1 & if\ x_k = y_k \\ \dfrac{1}{1+\log f_k(x_k) * \log f_k(y_k)} & otherwise \end{cases}$	Lower similarity for more frequent values mismatch. Higher similarity for less frequent values mismatch.
Goodall	$\begin{cases} 1-\sum_{q\in Q} p_k^2(q) & if\ x_k = y_k \\ 0 & otherwise \end{cases}$	Higher similarity to infrequent values. Lower similarity to frequent values.

2.2 Existing Algorithm

Squeezer algorithm [5],is a clustering algorithm for categorical data. It takes n tuples as input and produces clusters as output. Initially, the first tuple is read and cluster structure is constructed. Read subsequent tuples one after another. For each tuple, compute its similarities with all existing clusters. Select the largest similarity value. If the largest similarity value is greater than threshold's', the tuple is inserted into the existing cluster else new cluster is formed. The Cluster Structure (CS) will be updated for each iteration. Squeezer algorithm makes use of Cluster Structure which consists of cluster information and the summary information. Cluster summary includes information about the cluster, and normally it has pair of attribute values and its support value. Similarity measure used for measuring the similarity between cluster *C* and tuple *t* is given in equ 1.

$$Sim(C,tid) = \sum_{i=1}^{m} \left(\frac{\sup(a_i)}{\sum_j \sup(a_j)}\right) \quad (1)$$

Procedure

The steps involved in Squeezer algorithm are:

Step 1 Read the first tuple

Step 2 Generate the Cluster Structure (CS) .

Step 3 Read the next tuple and computes its similarity using support measure.

Step 4 If the similarity is greater than the threshold 's'add to the existing Cluster Structure. Else assign to the new Cluster Structure.

Step 5 Repeat Step 2 and 3 until the end of the tuple.

The advantages of Squeezer algorithm are as follows

- It produces high quality clustering result
- It deserves good scalability.
- It makes only one scan over the dataset, so it is highly efficient when considering I/O cost turn into bottleneck.

The disadvantages of Squeezer algorithm are as follows

- Impact of changing parameter's' on quality of cluster.

3. PROPOSED SYSTEM

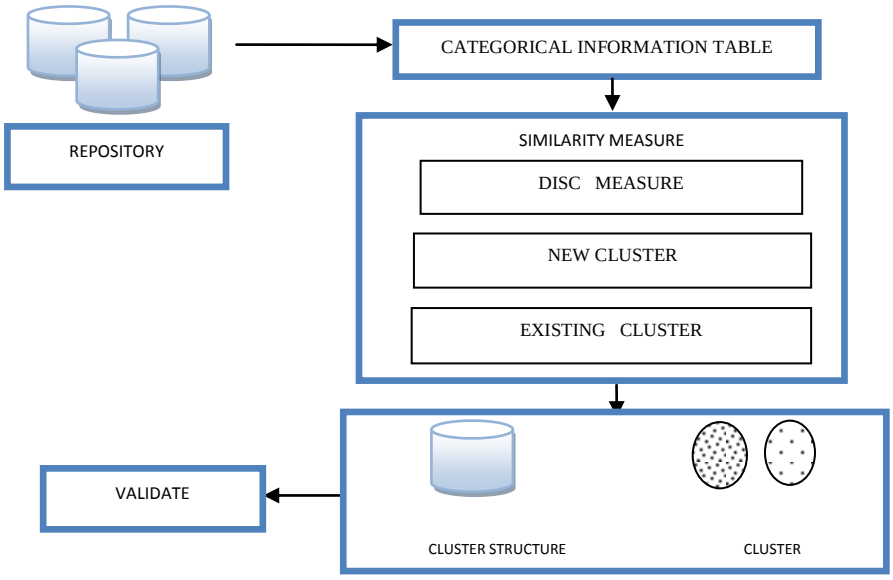

Fig 1: System Architecture of Enhanced Squeezer

System Architecture

The description of the proposed Enhanced Squeezer architecture is presented in fig 1.The main component of Enhanced Squeezer consist of

1. Categorical Information Table,
2. Similarity measure
3. Cluster Formation.
4. Cluster Validation

Categorical Information Table

Construct a data structure Categorical Information Table, which serves for the purpose of quick-reference for information related to co-occurrence statistics. Categorical Information Table consists of chance of occurrence of particular attribute value with that of another attribute value. Values in Categorical Information Table will be latter used for finding similarity between tuple using DISC measure. Consider the Balloon dataset that has 4 attributes and 16 instances. For a sample of two attributes namely color and size, the categorical Information Table that contains the co-occurrence
of color 'yellow' and 'purple' with that of the attribute values 'small' and 'large' is as shown below,

Table 1. Sample Categorical Information Table.

$Attribute_{(i)}$	$Attribute_{(i)}$	Co-Occurrence
Color:Yellow	Size: Small	1
Color:Yellow	Size: Large	1
Color:Purple	Size: Small	1
Color:Purple	Size: Large	1

3.2 Similarity Measure

In order to group data points into cluster, similarity between two data points must be computed. DISC (Data Intensive Similarity Measure for Categorical Data)[2] is used for measuring the similarity among data points. DISC identifies the semantics of the data without the help of domain experts for determining similarity. This is possible by capturing the relationships that are inherent in the data itself. DISC computes similarity between objects using cosine based similarity measure. Cosine similarity is the popular measure for document clustering. According to Enhanced Squeezer, the Similarity between categorical values v_{ij} and v_{ik} , Sim (v_{ij}, v_{ik}) is calculated by reading values from Categorical Information Table and apply it in following formula.

$$Sim=\sum_{v_{ml},v_{mk}\in A_m} \frac{CI[A_i:v_{ij}][A_m:v_{ml}]*CI[A_i:v_{ik}][A_m:v_{ml}]*sim(v_{ml},v_{ml})}{Normal\ Vector1*NormalVector2} \quad (3)$$

Where $CI[A_i: v_{ij}][A_m: v_{ml}]$ is the co-occurrence value obtained from categorical Information Table. $sim(V_{ml}, v_{ml}) = 1$ if $V_{ml} = v_{ml}$ and $sim(V_{ml}, v_{ml}) = 0$ if

$V_{mi} \neq v_{mi}$. If the similarity value found using the DISC measure is larger than threshold, then the tuple is assigned to existing cluster otherwise, constructs a New Cluster apart from the existing cluster. Based on the result of similarity computation, final cluster is constructed. Threshold value plays a vital role in clustering data. Here, threshold value is fixed by using steps as follows.

- Take 1/10 of dataset as sample.
- Find the similarity between the values.
- The average of similarity values is considered as threshold value‘s’.

3.3 Cluster Formation

Steps involved in Enhanced Squeezer algorithm
Step1.Read tuples iteratively and construct ‘Categorical Information Table’- (CI)
Step 3.Read the first tuple.
Step 4.Consider it as New Cluster
tep 5. For each subsequent tuple, compute similarity between tuple and existing cluster.

$$Sim=\sum_{v_{mi},v_{mi},A_m} \frac{CI[A_i=v_{ij}][A_m=v_{mi}]+CI[A_i=v_{ik}][A_m=v_{mj}]+sim(v_{mj},v_{mi})}{Normal\ Vector1+NormalVector2} \quad (3)$$

Step 6. Find the maximum value of similarity:
Step 7. If the maximum similarity is greater than the threshold value‘s’, Assign tuple to selected Existing Cluster, else construct New Cluster.
Step 8. Repeat Step 4-7, until end of the tuple.
Step 9. Extremely small clusters are discarded

3.4 Validation

The result of Enhanced Squeezer is evaluated to prove the degree of confidence of the results. Here, we compute the cluster accuracy and error rate to prove truthfulness of the proposed algorithm. To compute the cluster accuracy (r), we use the formula,

$$r = \frac{\sum_{i=1}^{k} a_i}{n} \quad (4)$$

where n denotes the number of instances in the dataset, a_i is the number of objects with class labels that dominates. Error rate (e) can be computed using the formula,

$$e = 1-r. \quad (5)$$

where ‘r’ represents the cluster accuracy.

3. EXPERIMENTAL STUDY

For our experiment, we use categorical dataset such as Balloon dataset, Mushroom dataset, Zoo dataset which taken from the UCI Repository. Let us now have a small introduction about these datasets.

Mushroom Dataset: Mushroom dataset is the most widely used dataset for testing the clustering algorithms. Mushroom dataset is broadly classified into classes as edible and poisonous. The Mushroom dataset has 22 attributes all are nominally valued and 8124 instances. Number of missing values in Mushroom dataset is 2480.

Zoo Dataset: Zoo dataset has 18 attributes with 101 instances. Class distribution of Zoo dataset has 7 classes. Missing value none.

Balloon Dataset: The first dataset is Balloon Dataset, which is fully categorical in nature. There are four dataset that represent the condition of the experiment. In the Balloon dataset, there are 4 attributes and the number of instances is 16.Missing value in Balloon dataset is nil.

Cluster Results

The outcome of Enhanced Squeezer is validated as discussed above in section 3.4. For instances, Consider Balloon dataset is grouped into 3 clusters. Each clusters with 8, 8, 4 objects respectively. 8 objects of cluster 1 can be categorized into 4 objects from class label 1, 4 objects from class label 2. 8 objects of cluster 2 can be categorized into 5 objects from class label 1, 3objects from class label 2. 4 objects of cluster 3 can be categorized into 4 objects from class label 1. As per the formula cluster accuracy for Balloon dataset is as follows $r=\frac{4+5+4}{20}$ = 0.65.Therefore, error e=1-0.65, e=0.35.

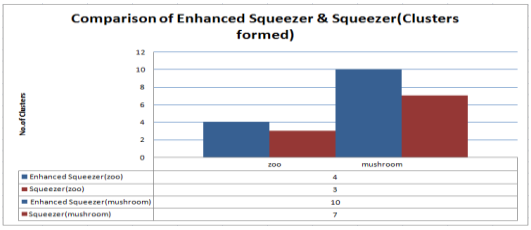

Fig 2: Comparison of Enhanced Squeezer & Squeezer based on No. Of Clusters.

The above figure represents the number of clusters formed for the existing system the proposed Enhanced Squeezer and is represented in bar chart. In this observation, the number of clusters formed for each of the datasets is higher in number than the Squeezer algorithm.

Experimental results of Enhanced Squeezer for Zoo and Mushroom dataset is given in the table.3 and table 4. The Cluster accuracy is more for the proposed work when compared with the normal squeezer algorithm. The result also shows that error rate is comparatively very small for the proposed work. From the observation of the tables that the DISC measure gives better results than the support measure for computing the similarity between categorical data**.** A dataset used here is slightly different from original format as in UCI machine learning repository

Table 3. Relative performance comparison of different Algorithm (Zoo dataset)

Algorithm	Cluster Accuracy	Error
Enhanced Squeezer	0.94	.06
Squeezer	0.72	0.28

Table 4. Relative performance comparison of different Algorithm (Mushroom dataset)

Algorithm	Cluster Accuracy	Error
Enhanced Squeezer	0.892	0.108
Squeezer	0.794	0.206

Table 5. Relative performance of algorithms based on Average error rate

Algorithm	Average Cluster Accuracy	Average Error
Enhanced Squeezer	0.916	0.084
Squeezer	0.757	0.243

4. CONCLUSION

Thus, Enhanced Squeezer algorithm using DISC measure is used for clustering categorical data. It involves thorough understanding of the domain and makes the data intensive. Cosine similarity is the popular measure for text clustering. DISC measure is based on cosine similarity measure. Enhanced Squeezer can handle large distinct attribute value efficiently using Categorical Information Table. Experimental results shows that clusters formed using proposed algorithm is more accurate and less error prone. Thus, Enhanced Squeezer can cluster categorical data effectively than existing algorithm and better algorithm to handle categorical data. Moreover, Enhanced Squeezer is generic and easy to implement.

5. REFERENCES

[1] Rishi Sayal and Vijay Kumar.V.2011. A novel Similarity Measure for Clustering Categorical Data Sets. International Journal of Computer Application (0975-8887).

[2] Aditya Desai, Himanshu Singh and Vikram Pudi. 2011. DISC Data-Intensive Similarity Measure for Categorical Data. Pacific-Asia Conferences on Knowledge Discovery Data Mining.

[3] Shyam Boriah, Varun Chandola and Vipin Kumar. 2008. Similarity Measure for Clustering Categorical Data. Comparative Evaluation. SIAM International Conference on Data Mining-SDM.

[4] Taoying Li, Yan Chen.2009. Fuzzy Clustering Ensemble Algorithm for partitional Categorical Data. IEEE, International conference on Business Intelligence and Financial Engineering.

[5] HE Zengyou, XU Xiaofei. 2002. SQUEEZER: An Efficient Algorithm for Clustering Categorical Data. Vol.17 No.5. Journal on Computer Science & Technology.

[6] D. Arthur and S. Vassilvitskii. 2007. K-means++: The advantages of careful seeding. In Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithm.

[7] Z.Haung and Michael K.Ng. 1999. A Fuzzy k-Modes Algorithm for Clustering Categorical Data. IEEE Transaction On Fuzzy systems, Vol. 7, No-4.

[8] Wang Jiacai and Gu Ruijun. 2010. An Extended Fuzzy K-Means Algorithm for Clustering Categorical Valued Data. International Conference on Artificial Intelligence and Computational Intelligence.

[9] Hua Yan, Keke Chen, Ling Liu3, Zhang Yil. 2009. SCALE: A Scalable Framework for Efficiently Clustering Transactional Data. Data mining and knowledge Discovery.

[10] Zengyou He, Xiaofei Xu,Shenchun Deng. 2008. k-ANMI: A Mutual Induction Based Clustering Algorithm for Categorical Data. Information Fusion9 (2).

[11] Zengyou He, Xiaofei Xu,Shenchun Deng . 2006. Improving Categorical Data Clustering Algorithm by Weighting Uncommon Attribute Value Matches. ComSIS Vol.3, No.1.

[12] Zengyou He, Xiaofei Xu. 2005. Scalable Algorithm for Clustering Large Dataset with Mixed Type Attributes. International Journal of Intelligent Systems.Vol.20.

[13] P.Gambaryan. 1964. A Mathematical Modeling of taxonomy. Izvest. Akad. Nauk Armen. SSR, 17(12).

[14] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. Second Edition Morgan Kaufmann publishers.