ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2014.0103

SURVEY ON CLUSTERING ALGORITHM AND SIMILARITY MEASURE FOR CATEGORICAL DATA

S. Anitha Elavarasi[1] and J. Akilandeswari[2]
[1]*Department of Computer Science and Engineering, Sona College of Technology, India*
E-mail: anishaer@gmail.com
[2]*Department of Information Technology, Sona College of Technology, India*
E-mail: akilandeswari@sonatech.ac.in

Abstract
Learning is the process of generating useful information from a huge volume of data. Learning can be either supervised learning (e.g. classification) or unsupervised learning (e.g. Clustering) Clustering is the process of grouping a set of physical objects into classes of similar object. Objects in real world consist of both numerical and categorical data. Categorical data are not analyzed as numerical data because of the absence of inherit ordering. This paper describes about ten different clustering algorithms, its methodology and the factors influencing its performance. Each algorithm is evaluated using real world datasets and its pro and cons are specified. The various similarity / dissimilarity measure applied to categorical data and its performance is also discussed. The time complexity defines the amount of time taken by an algorithm to perform the elementary operation. The time complexity of various algorithms are discussed and its performance on real world data such as mushroom, zoo, soya bean, cancer, vote, car and iris are measured. In this survey Cluster Accuracy and Error rate for four different clustering algorithm (K-modes, fuzzy K-modes, ROCK and Squeezer), two different similarity measure (DISC and Overlap) and DILCA applied for hierarchy and partition algorithm are evaluated.

Keywords:
Clustering, Categorical Data, Time Complexity, Similarity Measure, Data Mining Tools

1. INTRODUCTION

Data mining is a process of extracting useful information from the given data set. Data mining technique includes clustering, classification, regression, association, outlier detection etc. Clustering is a process of grouping objects with similar properties [1]. Clustering is an unsupervised learning. Any clustering process should exhibit high intra class similarity and low inter class similarity. Clustering algorithm can be broadly divided into hierarchical or partition algorithm. Hierarchical clustering algorithm group's data objects to form a tree shaped structure. It can be broadly classified into agglomerative hierarchical clustering (bottom up approach) and divisive hierarchical clustering (top down approach) Partition clustering algorithm splits the data points into k partition, where each partition represents a cluster. The partition is done based on certain objective function [2]. Similarity or dissimilarity measure of a clustering algorithm should exhibit the properties such as,

1. Symmetry : Sim(x, y) = Sim(y, x)
2. Non Negative: 0 < sim(x, y) < 1
3. Triangular Inequality : Sim(x, y) + Sim(y, z) = Sim(x, y)

Data in real world are either numerical or categorical in nature. Numerical data is continuous data and Categorical data consist of a set of categories. Categorical data are divided into Dichotomous and multi categorical data [23]. Dichotomous can have only two values. Multi categorical data can be in three ways, 1) an ordinal variable (ordered nature, e.g. high low medium), 2) nominal variable (unordered in nature, e.g. mode of transport preferred by persons) and 3) quantitative variable. Categorical data are used in health care, educational, marketing and biomedical field.

This paper describes about various clustering algorithm and similarity/dissimilarity measure applied to categorical data. This paper is organized as follows; section 2 gives an overview of different categorical clustering algorithms and its methodologies. Section 3 describes the time complexity of various categorical clustering algorithms. In section 4 various similarity measures used for categorical data are discussed. In section 5 the performance of various algorithm and similarity measure on the real world data sets are discussed. Finally in section 6, conclusions are provided.

2. EXISTING CATEGORICAL ALGORITHM

2.1 K MODES ALGORITHM

K means algorithm is a well known partition clustering algorithm. It is efficient for processing larger data set, sensitive to outliers and suitable only for numerical data set. The author [12] extends the k means by using simple matching dissimilarity function suitable for categorical data. Mode value is used instead of mean value and finally a frequency based method for updating the clustering process which reduces the cost function.

2.1.1 Methodology:

1. Choose *K* initial mode value.
2. Objective function used for categorical objects is,

$$d_c(X,Y)=\sum_{j=1}^{m}\delta(x_j,y_j) \tag{1}$$

and

$$\delta(x_j,y_j)=\begin{cases}0, & x_j=y_j\\ 1, & x_j\neq y_j\end{cases} \tag{2}$$

where, *X, Y* represents the categorical object and m refers to the categorical attribute.

3. Allocate an object to a cluster with minimum mode value. Update the mode for all iteration until end of the object.
4. Test the dissimilarity of object against current mode. If mode value of object belongs to different cluster rather

DOI: 10.21917/ijsc.2014.0103

than the current one, reallocate the object to the new cluster.

5. Repeat step 2, 3 and 4 until no such modification exists.

K modes algorithm produce only local optima. The author compares the performance and scalability of K-modes with K-prototype algorithm. Cluster performance is verified by using cluster accuracy and error rate for soya bean disease dataset. Soya bean has 47 instances with 35 attributes each. It can be classified under four diseases type. K modes algorithm is tested for soya bean dataset and produced 200 clusters with two different mode selections. A misclassification matrix is generated to analysis the cluster result with diseases classification. Scalability is verified against number of clusters for a given number of objects and number of objects for a given number of clusters using motor insurance dataset. Motor insurance has 690 instances described by 6 numerical and 9 categorical attributes with two possible classes. (Only 666 instances are used). K prototype algorithm produces 100 clusters. A misclassification matrix is generated to analysis the cluster with original classes.

2.2 SQUEEZER

Squeezer [9] [3] is a categorical data clustering algorithm. The main data structures involved are Cluster Summary and Cluster Structure. Summary holds set of pair of attribute value and their corresponding support value. Cluster Structure (CS) holds the cluster and summary information. The advantages of Squeezer algorithm are 1) It produces high quality cluster result 2) It deserves good scalability 3) It makes only one scan over the dataset, so it is highly efficient when considering I/O cost. The disadvantages of Squeezer algorithm is, the quality of the cluster depends on the threshold value(s). Space complexity is $O(n + k * p * m)$, where 'n' represent the size of the data set, 'm' represent number of attribute, 'k' represent final number of cluster and 'p' represent distinct attribute values.

2.2.1 Methodology:

1. Read the first tuple.
2. Generate the Cluster Structure (CS).
3. Read the next tuple and computes its similarity using support measure given as:

$$Sim(C,tid)=\sum_{i=1}^{m}\frac{\sup(a_i)}{\Sigma_j \sup(a_i)} \quad (3)$$

4. If the similarity is greater than the threshold 's'. Add to the existing Cluster Structure. Else assign to the new Cluster Structure.
5. Repeat Step 2 through 4 until the end of the tuple.

The author implements the algorithm in Java. It compares Squeezer algorithm with ROCK using Congressional vote dataset and Mushroom data set. Congressional vote dataset has 435 tuple with 16 attributes and 2 classes (democratic and republic). Mushroom has 8125 tuple with 22 attribute and 2 classes (poisonous and edible) Threshold values are assumed as 10 and 16 for vote and mushroom dataset respectively. The author concludes both algorithm produce high quality cluster. The only parameter that affects the clustering result and speed of the algorithm is threshold value(s).

2.3 ROCK

ROCK stands for **RO**bust **C**lustering using lin**K**s [4]. It is an Agglomerative hierarchy clustering. It uses links to measure similarity between data point. Initially each tuple is assigned as a separate cluster. Clusters are merged based on the closeness between clusters. Closeness is measured as the sum of the number of links between all pair of tuple. It is suitable for Boolean and categorical data. In traditional approach, categorical data are treated as Boolean value. Scalability of the algorithm depends on the sample size. The Criterion function and goodness measure used is given in Eq.(4) and Eq.(5).

Criterion function:

$$E_l=\sum_{i=1}^{k} n_i * \sum_{p_q,p_r\in c_i}\frac{link(p_q p_r)}{n_i^{1+2f(\theta)}} \quad (4)$$

where p_q, p_r represent the two points in a cluster and c_i represent the i^{th} cluster and n_i represent the size of the i^{th} cluster.

Goodness measure:

$$g(c_i c_j)=\frac{link[c_i c_j]}{(n_i+n_j)^{1+2f(\theta)}-n_i^{1+2f(\theta)}-n_j^{1+2f(\theta)}} \quad (5)$$

2.3.1 Methodology:

1. Draw a random sample
2. Compute the Link similarity
3. Cluster with the link
4. Label it on the disk

The author uses Congressional Vote dataset and Mushroom data set from UCI repository and compare ROCK algorithm with traditional centroid based hierarchical algorithm. Experiments were conducted on Sun Ultra-2/200 machine running Solaris 2.5 Operating system. In vote dataset cluster of republican contains only 12% of democrat whereas traditional approach has 25% of democrat with θ = 0.73. For Mushroom data set ROCK use θ = 0.8 and number of desired cluster as 20. It discovers pure clusters in the sense that mushroom in every cluster were either edible or poisons.

2.4 K-HISTOGRAM

K-Histogram extends k means algorithm to categorical domain by replacing mean with histogram and dynamically updates histogram during clustering process [13]. The K-means algorithm cannot cluster categorical data in an efficient way. To make them work for categorical data two modification is done. First mean value is replaced with histogram. Second new dissimilarity measure between categorical data and histogram is applied. Dissimilarity functions and cost measure applied for K histogram are given in Eq.(6) and Eq.(7).

Dissimilarity function used is:

$$d(H,Y)=\frac{\Sigma_{j=1}^{m} h_j y_j}{n} \quad (6)$$

Cost function used is,

$$P(W,H)=\sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l}\,d(X_i H_l) \quad (7)$$

Histogram can be used in computer vision application. Results of K histogram lie on the initial selection of Histogram and the order in which data are processed. Hence it produces only local optimal results.

2.4.1 Methodology:

1. Initialize the '*K*' value.
2. Apply cost function.
3. Allocate an object to a cluster whose histogram is near to it.
4. Update the histogram after each assignment.
5. Repeat the steps until no object change the cluster.

The author compares K Histogram with K modes algorithm for Congressional vote dataset and Mushroom data set. Algorithms were implemented in Java. Both K histogram and K modes uses same initial points selection method. Four comparisons were made, 1) Cluster error Vs Number of cluster, 2) Number of objects Vs Number of cluster, 3) Number of Iteration Vs Number of cluster, 4) pure cluster Vs Number of cluster.

2.5 ANALYSIS THE AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM FOR CATEGORICAL ATTRIBUTE

The author describes about the implementation detail of the K-pragna [11], an agglomerative hierarchical clustering algorithm. The Data structure used are Domain Array (DOM[m] [n]), Similarity Matrix and Cluster[m]. Domain Array holds the values of data set. Similarity matrix holds the similarity between the tuple / clusters. Cluster[m] is a single dimensional array holds the updated values whenever a merge occurs. The Language utilized is C.

2.5.1 Methodology:

1. Input the *k* (expected number of cluster) value.
2. Calculate the similarity.
3. Find the largest merge.
4. Repeat the step 2 and 3 till end.
5. Display the contents of each cluster.

The author used mushroom data set taken from UCI Machine Learning repository and tested the algorithm for $k = 3$. The accuracy of the algorithm is found to be 0.95.

2.6 HIERARCHICAL CLUSTERING ON FEATURE SELECTION FOR CATEGORICAL DATA OF BIOMEDICAL APPLICATION

The author [14] focuses on the feature association mining. Based on the contingency table, the distance (closeness) between features is calculated. Then hierarchical agglomerative clustering is applied. The clustered results helps the domain expects to identify the feature association of their own interest. The drawback of this system is it works only for categorical data.

2.7 FUZZY RULE BASED CLUSTERING ALGORITHM

Fuzzy Rule Based Clustering (FRBC) employs the supervised classification approach to do the unsupervised clustering [19]. It explores the potential clusters in data patterns and identifies them with fuzzy rules. Fuzzy clustering is applied when the cluster boundaries are vague. Advantages of fuzzy model is, it works with imprecise data, elements belong to more than one cluster with a specified degree of membership and the knowledge obtained are human readable FRBC are robust to noise and outlier.

2.7.1 Methodology:

1. Assume all unlabeled data patterns as Class 1.
2. Generate uniformly distributed instance as Auxiliary data and mark it as class2.
3. Apply SGERD (Steady state genetic algorithm to extract fuzzy Classification rule from data) rule generator to produce fuzzy rules to solve two class problem.
4. Select the best rule for class 1 and check whether it is less than the threshold, if less decrement the no of cluster and go to step 3. Else increment the cluster and remove from class1 and go to 3.

The author applies FRBC to 11 classification dataset and 2 clustering dataset obtained from UCI repository. FRBC is compared with other fuzzy clustering algorithm. The threshold (rule effectiveness measure) is set to 0.1 for all the dataset and the cluster specified by fuzzy rules are human understandable with good accuracy.

2.8 DBSCAN

DBSCAN stands for Density-based spatial clustering of applications with noise [24]. It is based on the notation of density reachablity. It requires two parameter, 1) Eps (Maximum radius of neighborhood) and 2) MinPts (Minimum number of points on the Eps neighborhood). Advantage of DBSCAN is it does not require the number of cluster prior, insensitive to the order of notation, and find arbitrarily shaped clusters. Drawback of this algorithm, quality depends on the distance function used.

2.8.1 Methodology:

1. Select a point *p*.
2. Retrieve all points from p satisfying Eps and MinPts.
3. If p is a core point, a cluster is formed.
4. If p is a border point and no points are density-reachable from, visits the next point of the database.
5. Repeat the process till all the points have been processed.

The author tests the efficiency with CLARANS using SEQUOIA dataset. DBSCAN is implemented in C++ based on R*Tree. Running time is compare for various numbers of points. DBSCAN outperforms CLARANS by a factor of more than 200.

DOI: 10.21917/ijsc.2014.0103

2.9 CURE (CLUSTERING USING REPRESENTATIVE)

CURE represents each cluster with a fixed number of points that are produced by selecting well scattered points from the cluster and then shrinking them towards the center of the cluster [20]. The scattered points after shrinking are the representatives for that cluster and then clusters with the closest pair of these representatives are merged repeatedly. It is an approach between the centroid-based and the all-point extremes. The time complexity of CURE is $O(s^2)$ for low-dimensional data where s is sample size of the data.

2.9.1 Methodology:

1. Draw random sample from the given data set.
2. Partition the sample.
3. Partially cluster the partitions.
4. Eliminate outliers.
5. Clusters the partial cluster.
6. Labeling data on a disk.

The parameter that affects the CURE algorithm are: shrinking factor (α), Number of representative points (c), sample size (s) and number of partition (p). The performance of CURE is compared with BRICH and Minimum Spanning Tree (MST). Results shows CURE can discover cluster with interesting shapes, less sensitive to outlier and less execution time is needed.

2.10 k-ANMI

The author [24] use the average normalized mutual information (Entropy based) as the criteria for the k-modes algorithm.

Objective function is defined as,

$$\phi^{ANIM}\left(\Lambda,\bar{\lambda}\right)=\frac{1}{r}\sum_{q=1}^{r}\phi^{NMI}\left(\bar{\lambda}\lambda^{(q)}\right) \quad (8)$$

Advantage of k-ANMI is, 1) suitable for both categorical data clustering and cluster ensemble, 2) it could be easily deployed in clustering distributed categorical data, 3) it is flexible in handling heterogeneous data that contains a mix of categorical and numerical attributes Limitation of k-ANMI is, 1) it is a great research challenge to implement the k- ANMI algorithm in an efficient way such that it is scalable to large datasets and 2) Finding global or near optimal is limited.

The author uses k-ANMI algorithm to Congressional vote dataset, Mushroom data set and Wisconsis Brest cancer data set from UCI repository. Cancer data set consist of 699 instance with 9 attributes and two class. Author compares k-ANMI algorithm with squeezer, GAClust, K-modes and ccdByEnsemble. k-ANMI outperforms all the other algorithm with respect to the average clustering error. Running time of kANMI algorithm increases linearly with number of object.

3. TIME COMPLEXITY

Time Complexity of any algorithm defines the amount of time taken by an algorithm to perform the elementary operation. Table.1 discusses the time complexity of various categorical clustering algorithm exist in the literatures [4, 9, 11, 12, 13, 18, 20, 21, 23, 24].

Table.1. Time Complexity of various clustering algorithm

Sl. No.	Algorithm	Time Complexity
1	K-modes	$O(tkn)$ t - No. of iteration k – No. of cluster n - No .of object
2	Squeezer	$O(n*k*p*m)$ n - Size of the data set k - Final number of cluster m - No. of attribute p - Distinct attribute values
3	ROCK	$O(n^2+nm_m m_a+n^2\log n)$ n - No. of input data point mm - Maximum no of neighbor ma - Average no of neighbor
4	K Histogram	$O(tkn)$ t – No. of iteration k – No. of cluster n – No. of object
5	Agglomerative hierarchical clustering Algorithm	$O(n^3)$ n – No. of objects
6	Genetic K Means	$O(nd)$ - fitness function $O(n^2d)$ - mutation $O(nKd)$ - K-means n - Size of the data set k - Final number of cluster d - Dimensions of data set
7	K-ANMI	$O(Ink^2rp^2)$ n - Size of dataset r - Number of attributes k - Number of the histograms, the size of every histogram, the number of clusters I - iteration times p - number of distinct attributes values
8	CURE	$O(n^2 \log n)$ n - Input Size
9	DBSCAN	$O(m*\log(m))$ m – No. of points in database
10	CLOPE	$O(N*K*A)$ N -total number of transaction K - No of Cluster A - average length of the transaction

4. EXISTING CATEGORICAL SIMILARITY MEASURES

4.1 Chi - SQUARED

Karl Pearson in 1900 proposes the chi-squared Statistic [15]. It examines whether there exist, any association between the categorical variable. Range exists between -1 to +1 for two variables and 0 to +1 for larger number of variable. The value more close to 1 indicates a strong relationship between variables. The chi square (x^2) formula is defined as,

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{9}$$

where, O_i represent observed value and E_i represent Expected value.

Steps in Chi Square Test:

1. Given Observed frequency
2. Note the Expected frequency
3. Apply the chi square formula
4. Find the degree of freedom($df = N - 1$)
5. If the obtained value is equal or greater than the chi square table reject the null hypothesis.

Advantage of Chi square is it requires no assumptions about the shape of the population distribution from which a sample is drawn. It can be applied to nominal or ordinal measured variables. Limitation of Chi square similarity are, 1) need quantitative data, 2) sensitive to sample size, 3) does not give much information about the strength of the relationship and 4) Expected frequency should not be less than 1.

4.2 COSINE SIMILARITY

Cosine similarity [17] is a popular method for text mining. It is used for comparing the document (word frequency) and finds the closeness among the data points in clustering. Its range lies between 0 and 1. The similarity between two terms X and Y are defined as follows.

$$CosineSim(X,Y) = \frac{X^{o}Y}{\|X\|\|Y\|}. \tag{10}$$

One desirable property of Cosine similarity is independent of document length. Limitation is the terms are assumed to be orthogonal in space. If the value is zero no similarity exist between the data element and if the vale is 1 similarity exist between two elements. Considered two documents X and Y with attributes X = {1 2 3 0 0} and Y = {2 4 0 0 1},

$$CosineSim(X,Y) = \frac{1*2+2*4+3*0+0+0*1}{\|1*1+2*2+3*3\|\|2*2+4*4+1*1\|} = 0.5832$$

4.3 OVERLAP

The overlap measure counts the number of attribute that matches the two data instance. It uses only the diagonal entries of the similarity matrix and sets off diagonal entries to 0 [5]. The range of per attribute value is 0 to 1. 0 indicate no match exist between the attribute and 1 indicates match exist between the attribute. The overlap similarity is defined as,

$$S_k(X_k, Y_k) = \begin{cases} 1 & \text{if } X_k = Y_k \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

4.4 DISC

Data Intensive Similarity Measure for Categorical Data analysis (DISC) [6]. It makes use of a data structure called categorical information table (CI Table). CI table stores the co-occurrence statistics for the categorical data. The similarity between two attribute is measured using the cosine similarity measure.

4.4.1 Methodology:

1. Construct the Categorical Information table (CI Table)
2. Initialization of similarity matrix.

$$\forall i, j, k : sim(v_{ik}, v_{jk}) = 0 \text{ if } v_{ik} \neq v_{jk} \tag{12}$$

$$\forall i, j, k : sim(v_{ik}, v_{jk}) = 1 \text{ if } v_{ik} \neq v_{jk} \tag{13}$$

3. Computer the Similarity between two attribute(v_{ij},v_{ik}) using the

$$Sim(v_{ij}, v_{ik}) \frac{1}{d-1} \sum_{m=1, m\neq i}^{d} Similarity_m \tag{14}$$

Similarity$_m$ = Cosine Product ($CI[A_i:v_{ij}][A_m]$, $CI[A_i:v_{ik}][A_m]$) for Categorical data

$$Similarity_m = 1 - \frac{CI[A_m : v_{ij}][A_m] CI[A_j : v_{ik}][A_m]}{\max[A_m]\min[A_m]} \tag{15}$$

for Numerical data where the cosine product is defined as,

$$Sim = CS = \sum_{v_{ml}, v_{ml} \in A_m} \frac{CI[A_i : v_{ij}][A_m : v_{ml}] * CI[A_i : v_{ik}][A_m : v_{ml}] * sim(V_{ml}, v_{ml})}{NormalVector1 * NormalVector2} \tag{16}$$

4. Repeat the step 2 and step 3.

The author concludes that DISC outperforms other similarity measure both for classification and regression analysis.

4.5 DILCA

DILCA - DIstance Learning in Categorical Attribute is the measure used by the author [7, 8]. Co-occurrence table is formed for all the features using symmetric uncertainty a matrix is generated and conditional probability is applied, the results are given to the Euclidean measure to find the similarity between the attributes.

4.5.1 Methodology:

1. Context Selection (Feature extraction) is based on symmetric uncertainty (SU). It is a co-relation based measure from information theory. The co-relation matrix are formed using SU,

$$SU(X,Y) = 2 * \frac{IG(X/Y)}{H(X) + H(Y)} \tag{17}$$

DOI: 10.21917/ijsc.2014.0103

where, $IG(X/Y)$ is the information gain and $H(X)$ and $H(Y)$ represent the entropy of the variable X and Y respectively.

2. Distance Computation: Applying Conditional probability for the co-relation matrix and Euclidean distance,

$$d(x_i x_j) = \sqrt{\sum_{Y \in content(X)} \sum_{y_k \in Y} \left(P(x_i / y_k) - P(x_j / y_k)\right)^2} \quad (18)$$

The author embedded the similarity measure both on partition and hierarchical algorithm. The results are scalable with respect to the number of instance in the dataset.

5. PERFORMANCE ANALYSIS

The cluster validation is the process of evaluating the cluster results in a quantitative and objective manner. Cluster evaluation is done either internal or external. The internal evaluation determines the quality of the cluster. The external evaluation determines the partitioning among the cluster. The results of different clustering algorithm are validated using Cluster accuracy, error rate. Cluster Accuracy 'r' is defined as,

$$r = \frac{\Sigma_{i=1}^{k} a_i}{n} \quad (19)$$

where, 'n' refers number of instance in the dataset, 'a_i' refers to number of instance occurring in both cluster i and its corresponding class and 'k' refers to final number of cluster.

Error rate 'E' is defined as, $E = 1 - r$, where 'r' refers to the cluster accuracy.

Real world dataset: Five real life dataset, such as Mushroom, vote, Iris, cancer and zoo obtained from UCI machine learning repository [25]. **Mushroom:** Each tuple represent the physical characteristic of mushroom. Number of instance is 8124 and number of attribute is 22. It can be classified into edible (4028) and poisonous mushroom (3916). **Vote:** Each tuple represent the United States congressional vote record in 1984. Number of instance is 435 and number of attribute is 16. It can be classified into Democrat (267) and Republican (168). **Iris:** The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. Number of instance is 150 and number of attribute is 4. **Cancer:** No of Instance 8124 and no of attribute are 22. **Zoo:** Zoo dataset has 18 attributes with 101 instances. Class distribution of Zoo dataset has 7 classes. **Soyabean:** Number of instance is 307 and number of attribute is 35.

In this survey Cluster Accuracy and Error rate are evaluate in three ways, 1) comparisons of different categorical clustering algorithm, 2) comparisons of DILCA combined with partition and hierarchical clustering algorithm and 3) comparisons of categorical similarity measure.

5.1 COMPARISONS OF DIFFERENT CATEGORICAL CLUSTERING ALGORITHM

Algorithm used for comparisons are K-modes [10] [12] [16] and fuzzy K-modes [16] using soya bean and Zoo data set is depicted in Fig.1. Algorithm was run for 100 times and fuzzy parameter is set as 1.1. The author [16] makes use of four real life dataset to show the accuracy, precession and recall values. To illustrate the efficiency of the algorithm synthetic datasets are generated and four different graphs is plotted for, 1) Time Vs number of cluster, 2) Time Vs Number of Objects, 3) Time Vs Number of categories and 4) Time Vs number of Attributes. Fuzzy K-modes outperforms K -modes for both the data set.

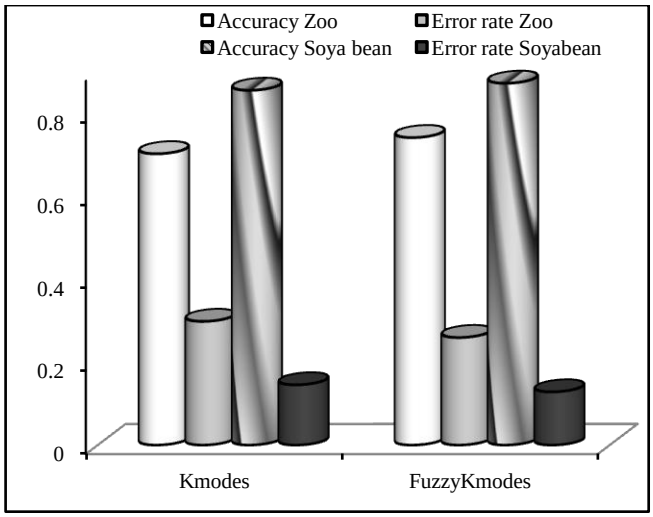

Fig.1. Comparison of K modes and Fuzzy K modes

ROCK and Squeezer [9] for mushroom data set is depicted in Fig.2. Squeezer outperforms ROCK for the mushroom data set and ROCK outperform Squeezer for the vote data set. The author [8] compares ROCK with hierarchical and partitioned algorithm for four real world data sets. Threshold values for ROCK are set as 0.2 to1 in step of 0.05.

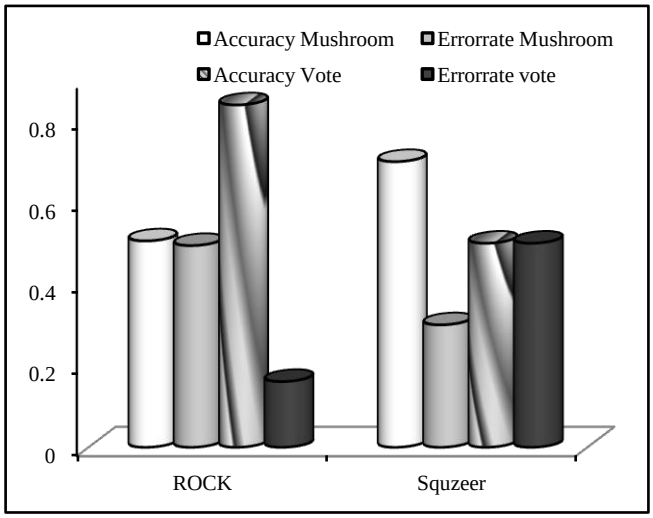

Fig.2. Comparison of Rock and Squeezer

5.2 COMPARISONS OF DILCA COMBINED WITH PARTITION AND HIERARCHICAL CLUSTERING ALGORITHM

The author [8] determines the quality of cluster formed using accuracy and normalized mutual information. The partition (K-modes algorithm) and wards hierarchical clustering algorithm are combined with DILCA for mushroom, vote and cancer dataset are depicted in Fig.3. Both the algorithm set the number

of cluster equal to the number of classes. DILCA_Kmodes algorithm is implemented using WEKA platform and DILCA_Hierarchical algorithm is implemented using Java Murtagh's platform.

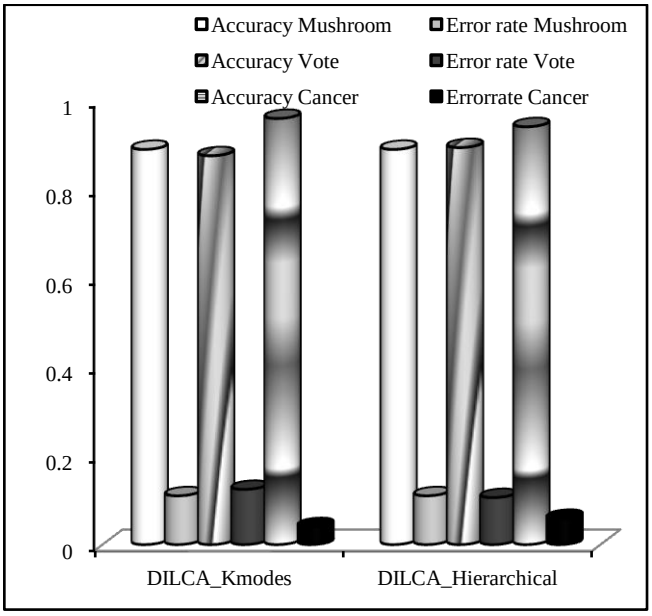

Fig.3. Comparison of DILCA for parametric and nonparametric

5.3 COMPARISONS OF CATEGORICAL SIMILARITY MEASURE

The author [6] uses 24 different dataset and 15 similarity measures were assessed for classification and regression using the kNN algorithm. Experiments were conducted using WEKA environment. Results were presented for 10 fold cross validation The categorical similarity measures used for comparisons in fig.4 are DISC and overlap for the car evaluation, iris and cancer dataset using kNN algorithm with K = 10. Both similarity measures give similar results for all the three dataset.

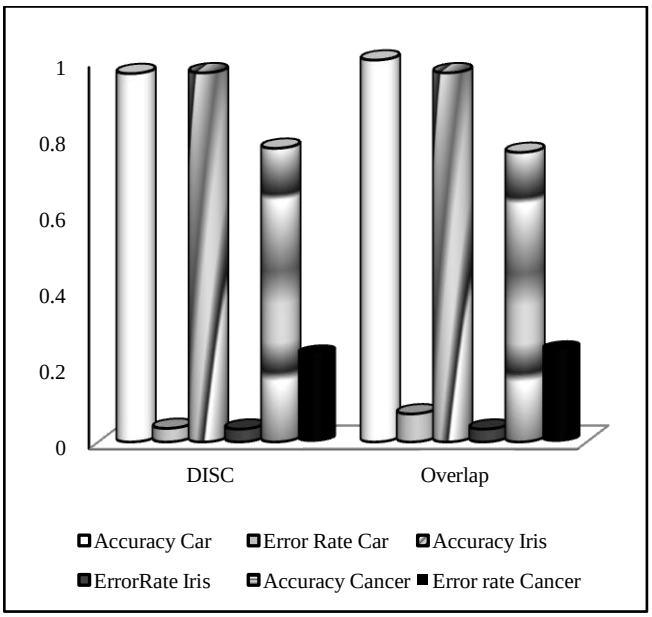

Fig.4. Comparison of DISC and Overlap

Several open source data mining tools are available on web. Table.2. illustrates various data mining tools and few clustering algorithm available.

Table.2. Data mining tool and clustering algorithm

Tool Name	Clustering Algorithm
Weka	K- Means, X-Means, EM, Cobweb CLOPE, OPTICS, DBSCAN Hierarchical clustering algorithm
R package	K -Means, PAM, DBSCAN, ROCK Hierarchical clustering algorithm
Rapidminer	K -Means, DBSCAN EM K-Medoids X-means Kernel K- means Fast K - Means Hierarchical clustering algorithm

6. CONCLUSION

The paper describes a review on different clustering methodologies and similarity measure associated with the categorical data clustering. The factor that affects various clustering algorithm, its advantage and limitation are discussed. Time complexities of various categorical clustering algorithms are discussed. Cluster accuracy and error rate for real world data set using different categorical clustering algorithms, parametric and non parametric version of DILCA and categorical similarity measure are illustrated.

REFERENCES

[1] J. Han and M. Kamber, "Data Mining Concepts and Techniques", *The Morgan Kaufmann Series in Data Management Systems*, Morgan Kaufmann, 2000.

[2] S. A. Elavarasi, J. Akilandeswari and B. Sathiyabhama, "A Survey on Partition Clustering Algorithms", *International Journal of Enterprise Computing and Business Systems*, Vol. 1, No. 1, pp. 1-14, 2011.

[3] R. Ranjani, S. A. Elavarasi and J. Akilandeswari, "Categorical Data Clustering using Cosine based similarity for Enhancing the Accuracy of Squeezer Algorithm", *International Journal of Computer Applications*, Vol. 45, No. 20, pp. 41-45, 2012.

[4] S. Guha, R. Rastogi and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes", *Information Systems*, Vol. 25, No. 5, pp. 345 – 366, 2000.

[5] S. Boriah, V. Chandola and V. Kumar, "Similarity Measures for Categorical Data: A Comparative Evaluation", *Proceedings of the 8th SIAM International Conference on Data Mining*, pp. 243 – 254, 2008.

[6] A. Desai, H. Singh and V. Pudi, "DISC: Data Intensive Similarity Measure for Categorical Data", *Proceedings of Advances in Knowledge Discovery and Data Mining – 15th Pacific Asia Conference*, Vol. 6635, pp. 469 – 481, 2011.

[7] D. Ienco, R. G. Pensa and R. Meo, "From Context to Distance: Learning Dissimilarity for Categorical Data Clustering", *ACM Transactions on Knowledge Discovery from Data*, Vol. 6, No. 1, 2012.

[8] D. Ienco, R. G. Pensa and R. Meo, "Context-based distance learning for categorical data clustering", *Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII*, pp. 83 – 94, 2009.

[9] H. Zengyou, X. Xiaofei and D. Shengchun "Squeezer: An Efficient Algorithm for Clustering Categorical Data", *Journal on Computer Science & Technology*, Vol. 17, No. 5, pp. 611-624, 2002.

[10] Z. Haung and M. K. Ng, "A Fuzzy k-Modes Algorithm for Clustering Categorical Data", *IEEE Transactions on Fuzzy systems*, Vol. 7, No. 4, pp. 446-452, 1999.

[11] P. Agarwal, M. Afshar Alam and R. Biswas, "Analysis the Agglomerative hierarchical clustering Algorithm for Categorical Attribute", *International Journal of Innovation, Management and Technology*, Vol. 1, No. 2, pp. 186 – 190, 2010.

[12] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", *Data Mining and Knowledge Discovery*, Vol. 2, No. 3, pp. 283 – 304, 1998.

[13] Z. He, X. Xu, S. Deng and B. Dong, "K- Histograms: An Efficient Clustering Algorithm for Categorical Dataset", *CoRR*, Vol. abs/cs/0509033, 2005.

[14] Y. Lu and L. R. Liang, "Hierarchical Clustering of Features on Categorical Data for Biomedical Applications", *Proceedings of the ISCA 21st International conference on Computer Applications in Industry and Engineering*, pp. 26 - 31, 2008.

[15] Alan Agresti, "*An Introduction to categorical data analysis*", Wiley Series in Probability and Statistics, Second Edition, Wiley-Interscience, 2007.

[16] Michael K. Ng and Liping Jing, "A new fuzzy k-modes clustering algorithm for categorical data", *International Journal on Granular Computing, Rough Sets and Intelligent Systems*, Vol. 1, No. 1, pp. 105 – 119, 2009.

[17] A. Rajaraman and J. D. Ullman, "*Mining of Massive Datasets*", Cambridge University Press, 2011.

[18] D. K. Roy and L. K. Sharma, "Genetic K means Clustering Algorithm for Mixed Numerical and Categorical data set", *International journal of Artificial Intelligence & Applications*, Vol. 1, No. 2, pp. 23 – 28, 2010.

[19] E. G. Mansoori, "FRBC: Fuzzy Rule Based Clustering Algorithm", *IEEE Transactions on Fuzzy Systems*, Vol. 19, No. 5, pp. 960 – 971, 2011.

[20] S. Guha, R. Rastogi and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 73 - 84, 1998.

[21] Y. Yang, X. Guan and J. You, "CLOPE: a fast and effective clustering algorithm for transactional data", *Proceedings of the 8th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 682 – 687, 2002.

[22] H. Rezankova, "Cluster Analysis and categorical data" *Statistika*, pp. 216 - 232, 2009.

[23] M. Ester, H. P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases", *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 222 – 231, 1996.

[24] H. Zengyo, X. Xu and S. Deng, "k-ANMI: A mutual information based clustering algorithm for categorical data", *Information Fusion*, Vol. 9, No. 2, pp. 223 – 233, 2008.

[25] UCI Machine Learning Repository, http://www.ics.uci.edu/mlearn/MLRepository.html.