Research Journal of Applied Sciences, Engineering and Technology 11(7): 798-805, 2015
DOI: 10.19026/rjaset.11.2043
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: June 8, 2015 Accepted: July 8, 2015 Published: November 05, 2015

Research Article

Term Frequency Based Cosine Similarity Measure for Clustering Categorical Data using Hierarchical Algorithm

S. Anitha Elavarasi and J. Akilandeswari
Department of Computer Science and Engineering, Sona College of Technology, Salem, Tamil Nadu, India

Abstract: Object in real world are categorical in nature. Categorical data are not analyzed as numerical data because of the absence of inherit ordering. In this study performance of cosine based hierarchical clustering algorithm for categorical data is evaluated. It make use of two functions such as Frequency Computation, Term Frequency based Cosine Similarity Matrix (TFCSM) computation. Clusters are formed using TFCSM based hierarchical clustering algorithm. Results are evaluated for vote real life data set using TFCSM based hierarchical clustering and standard hierarchical clustering algorithm using single link, complete link and average link method.

Keywords: Categorical data, clustering, cosine similarity, hierarchical clustering, term frequency

INTRODUCTION

Data mining deals with extracting information from a data source and transform it into a valuable knowledge for further use. Clustering is one of the techniques in data mining. Clustering deals with grouping object, which are similar to each other. Clustering process should exhibit high intra class similarity and low inter class similarity (Jiawei *et al*., 2006). Clustering algorithms are broadly classified into partition algorithms and hierarchical algorithms.

Hierarchical clustering algorithms group data objects to form a tree shaped structure. It can be broadly classified into agglomerative hierarchical clustering and divisive hierarchical clustering. Agglomerative approach is also called as bottom up approach, where each data points are considered a separate cluster. In each iteration, clusters are merged based on certain criteria. The merging can be done by using single link, complete link and centroid or wards method. Divisive approach otherwise called as top down approach, where all data points considered as a single cluster and they are split into number of clusters based on certain criteria. Advantages of this algorithm are:

- No priori information about the number of cluster is required
- Easy to implement.

Drawbacks of this algorithm are:

- Algorithm can never undo what was done previously
- Sensitivity to noise and outliers
- Difficult to handle convex shapes

Time complexity is $O(n^2 \log n)$ where n is the number of data points. Examples for hierarchical clustering algorithms are LEGCLUST (Santos *et al*., 2008), BRICH (Balance Iterative Reducing and Clustering using Hierarchies) (Virpioja, 2008), CURE (Cluster Using REpresentatives) (Guha *et al*., 1998).

Objects in real world are categorical in nature. Categorical data is not analyzed as numerical data because of the absence of implicit ordering. Categorical data consists of a set of categories as a dimension for an attribute (Agresti, 1996; Agresti, 2013). Categorical variables are of two types. They are:

- Ordinal variable (variables with ordering e.g.: patient condition can be expressed as good, serious and critical.)
- Nominal variable (variables without a natural ordering e.g.: type of music can be folk, classical, western, jazz, etc)

Cosine similarity (Jiawei *et al*., 2006) is a popular method for information retrieval or text mining. It is used for comparing the document (word frequency) and finds the closeness among the data points. Distance or similarity measure plays vital role in the formation of final clusters. Distance measure should satisfy three main properties such as:

Corresponding Author: S. Anitha Elavarasi, Department of Computer Science and Engineering, Sona College of Technology, Salem, Tamil Nadu, India

- Non negativity
- Symmetry
- Triangular inequality

Popular distance measures are euclidean distance and manhattan distance. In this study term frequency based cosine similarity has been applied to all the three versions (single, average and complete linkage) of hierarchical clustering algorithms. Performance of term frequency based cosine and standard cosine similarity for hierarchical clustering algorithms are analyzed.

LITERATURE REVIEW

LEGclust is a hierarchical agglomerative clustering algorithm based on Renyi's Quadratic Entropy (Santos *et al*., 2008). For a given set of data points X = {x1, x2…xn}, each element of the dissimilarity matrix A is computed by using Renyi's Quadratic entropy. A new proximity matrix L is built by using dissimilarity matrix. Each column of the proximity matrix corresponds to one layer of connections. By use this proximity matrix, the subgraphs for each layer are built. K minimum number of connections is defined. Clusters with maximum number of connections are merged on each iteration The parameter involved in clustering process are number of nearest neighbors, smoothing parameter and minimum number of connections used to join cluster in each iteration. Experiments were conducted both on real life data set (Olive, Wine, 20NewsGroups, etc) as well as synthetic datasets. Results indicate that LEGClust achieves good results, simple to use and valid for datasets with any number of features.

CURE (Clustering Using Representatives) (Guha *et al*., 1998) is more efficient in the presence of outliers and identifies clusters with non-spherical shapes. It 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 center of the cluster. The scattered points after shrinking are chosen as representatives for that cluster. The closest pair of these representatives is merged repeatedly to form the final clusters. It is an approach between the centroid-based and the all-point extremes. CURE algorithm is sensitive to shrinking factor, number of representative points, number of partition and random sample size. The time complexity of CURE is $O(s^2)$ and space complexity is O(s) for low-dimensional data, where s is sample size of the data set. Advantages of CURE are:

- Less sensitive to outlier
- Low execution time

BIRCH **(**Balanced Iterative Reducing and Clustering using Hierarchies) (Virpioja, 2008) identifies clusters with the available resources such as limited memory and time constraints. It can find good clusters with a single scan of the data and quality of cluster is improved with additional scans. Hence I\O cost is linear. BRICH make use of CF (Clustering Feature) tree. CF is a triplet consisting of <N, LS and SS>, where N refers to the number of data points in the cluster, LS refers to the linear sum of the N data points and SS refers to the square sum of the N data points. BRICH algorithm is sensitive to initial threshold, page size, outlier and memory size. Scalability of BRICH is tested by increasing the number of points per cluster and increasing the number of cluster. BRICH is more accurate, less order sensitive and faster.

Analysis of the Agglomerative hierarchical clustering Algorithm for Categorical Attribute describes about the implementation detail of the K-pragna (Agarwal *et al*., 2010), an agglomerative hierarchical clustering algorithm. Data structures 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 which holds the updated values whenever a merge occurs. Expected number of cluster given as input. Similarity is calculated among instances. Clusters are formed by merging the data points. 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.

Hierarchical clustering on feature selection for categorical data of biomedical application (Lu and Liang, 2008) focuses on the feature association mining. Based on the contingency table, the distance (closeness) between features is calculated. Hierarchical agglomerative clustering is then applied. The clustered results helps the domain expects to identify the feature association of their own interest. The drawback of this system is that it works only for categorical data.

CHAMELEON (Karypis *et al*., 1999) measures the similarity of two clusters based on a dynamic model. It does not depend on a static model supplied by the user since it considers both the natural characteristics of the cluster such as Relative Interconnectivity and Relative Closeness. The relative inter connectivity between clusters is defined as the absolute interconnectivity between them and normalized with respect to the internal inter-connectivity of those clusters. The relative closeness between clusters is defined as the absolute closeness between them is normalized with respect to the internal closeness of those clusters. Sparse graph constructed for the given set of data points, where each node represents data items and weighted edge represents similarities among the data items. Cluster the data items into a large number of small sub-clusters using a graph partitioning algorithm. It finds the clusters by repeatedly combining these sub-clusters using an agglomerative hierarchical clustering algorithm.

Cosine similarity (Jiawei *et al*., 2006) is a popular method for information retrieval or text mining. It is

Table 1: Clustering algorithm for categorical data

Algorithm	Methodology	Complexity	Advantage	Disadvantage
ROCK (Guha *et al*., 1999)	Clustering categorical data is done by neighbor and link analysis.	$O(n^2+nm_m m_a+n^2 \log n)$ n- Number of input data point m_m-Maximum number of neighbor. ma- Average number of neighbor.	Generate better quality clusters and exhibits good scalability.	Closeness among elements of cluster is missing.
Squeezer (He *et al*., 2002)	Reads each tuple in sequence assigning the tuple to an existing cluster or creating a new cluster determined by the similarities between tuple and clusters.	O(n*k*p*m) n- Size of the data set. k- Final number of cluster. m- Number of attribute. p- Distinct attribute values.	It makes only one scan over the dataset	The quality of the cluster depends on the threshold value.
BIRCH	Integration of hierarchical clustering with the iterative partition method. It uses CF tree to dense leaf node into a separate cluster.	O(n) n-number of objects.	Handle Noise and does single scan of data set.	Sensitive to initial threshold, outliers and memory size.
CURE	It 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 center of the cluster.	$O(n^2)$ n - Number of sample size.	Handle outliers.	Sensitive to shrinking factor, number of representative points and number of partition.
LEGClust	Agglomerative hierarchical clustering algorithm based on Renyi's Quadratic Entropy.	$O(N((M/2)+1)^2)$ N- Number of objects. M-number of nearest neighbor.	Achieves good results and simple to use.	Sensitive to smoothing parameter (h) and number of nearest neighbor (M).

used for comparing the document (word frequency) and finds the closeness among the data points during clustering. Its range lies between 0 and 1. The similarity between two terms X and Y are defined as follows:

$$CosineSim(X,Y) = \frac{X^{\circ}Y}{\|X\|\|Y\|} \quad (1)$$

One desirable property of cosine similarity is that it is independent of document length. Limitation of the method is that the terms are assumed to be orthogonal in space. If the value is zero, no similarity exists between the data elements and if the vale is 1 similarity exists between two elements.

Table 1 gives the overall comparison of various clustering algorithm used for categorical data. The methodology of various algorithms, its complexity, pros and cons are summarized.

Term frequency based cosine similarity for hierarchical clustering algorithm: In this study term frequency based cosine similarity measure has been used for clustering categorical data. The most popular hierarchical clustering algorithm is chosen as an underlying algorithm. The core part of this methodology is similarity matrix formation. Data from real word consist of noise and inconsistency. Data preprocessing ensures quality input to be given to the similarity computation process.

Similarity matrix formation uses two functions:

Term frequency computation: Term Frequency computation deals with calculating the rate of occurrence for each attributes available in the dataset.

Term Frequency based Cosine Similarity (TFCS): TFCS deals with computing the similarity matrix using cosine similarity defined in Eq. (2). Frequencies are generated and stored in a multi dimensional array. Similarity matrix generated is given as an input for the hierarchical clustering algorithm. Clusters are formed and the results are evaluated.

Definition: This section describes the definitions used in this model. Let Z be the data set with X_1 to X_n instance. Each instance has $A_1..A_n$ categorical attributes with D_1 to D_n domain respectively. Val(D_{ij}) finds the number of times a particular value occur in a domain. TSim[X, Y] represents a similarity matrix computed using TFCS.

Definition 1: Frequency computation deals with calculating the rate of occurrence for each attribute present in the dataset. In other words it returns val(Dij) (i.e., number of times a particular value occur in a domain Di for an attribute Ai). It is represented by the term F_w.

Definition 2: Term Frequency based Cosine Similarity computes the similarity matrix TSim[x,y] for the given data D. Let X and Y be the two instances with n attribute. TFCS(X, Y) is defined as:

$$TFCS(X,Y) = \frac{\sum_{i=1}^{n} F(X_i)*F(Y_i)}{\sqrt{\sum_{i=1}^{n}(F(X_i))^2}*\sqrt{\sum_{i=1}^{n}(F(Y_i))^2}} \quad (2)$$

Definition 3: Let X and Y are the two instances of the given data D. Cosine similarity computes the similarity between X,Y and is defined as:

Table 2: Car data set

ID	Buying	Maintenance	Door	Person
A	Vhigh	Med	Two	Four
B	Vhigh	Med	Two	Four
C	Vhigh	Med	Two	More
D	Vhigh	Med	Three	Four
E	Vhigh	Med	Three	More
F	Vhigh	Small	Two	Four
G	Vhigh	Small	Two	Four
H	Vhigh	Small	Two	More
I	Vhigh	Small	Three	Four
J	Vhigh	Small	Three	More

Table 3: Similarity computation

Similarity between ID	Cosine similarity	TF based cosine similarity
A, B	1.00	1.00
A, C	0.75	0.86
A, D	0.75	0.86
A, E	0.50	0.71
A, F	0.75	0.86
A, G	0.75	0.86
A, H	0.50	0.71
A, I	0.50	0.71
A, J	0.25	0.56

$$CS(X,Y) = \frac{\sum_{i=1}^{n} X_i * Y_i}{\sqrt{\sum_{i=1}^{n}(X_i)^2} * \sqrt{\sum_{i=1}^{n}(Y_i)^2}} \quad (3)$$

Algorithm:

Input: Given Dataset with 'N' categorical attributes and threshold
Output: 'k' clusters
Algorithm: TFCSHCA
begin
read the dataset one by one
Initialize no of cluster c
//Computer frequency
Assign initial occurrence of an attribute A_i of domain D_{ij} be zero
while not EOF do
    For each attribute A_i of domain D_{ij}
      Calculate the occurrence of each attribute as Fw_{ij}
    End for
  End while
//Similarity Matrix formation using frequency count
Vector representation of string (X,Y)
Multiplying each vector by its frequency of occurrence Fw
For i = 1 to vectorlength do
    XYvector = $F_w(X_i)$ * $F_w(Y_i)$
    Xvector = $F_w(X_i)$ * $F_w(X_i)$
    Yvector = $F_w(Y_i)$ * $F_w(Y_i)$
    compute TSim(X,Y) using Eq. (2)
End for
//Cluster the data with max similarity-HCA
Initialize each cluster to be a singleton.
Fix the initial threshold 't'
While (TSim (X, Y) <= t)
Begin
  Find the two closest clusters using similarity matrix.
  Merge the closest cluster.
  Assign to the respective cluster Cluster [c++]
  Update the similarity matrix
End of while loop
Return final cluster formed
End

Sample walkthrough: Let us consider the Car dataset [uci] shown in Table 2, with 10 number of instance (A to J) and 4 attributes. Attribute information used in balloon data set are color, size, act and age. Domain of Buying = 1 (i.e., Vhigh), Domain of maintenance = 2 (i.e., small and med), Domain of door = 2 (i.e., two and three) and Domain of person = 2 (i.e., four and more).

The term frequencies for each element are represented in the array F. The frequency for each attribute are: F[Vhigh] = 10, F[Small] = 5, F[Med] = 5, F[Two] = 6, F[Three] = 4, F[Four] = 6 and F[More] = 4. Similarity computation of A with all other element is represented in Table 2. The detailed computation of similarity for (A, C) and (A, E) for both cosine similarity and term frequency based cosine similarity are shown below. Table 3 represents the similarity computation for car datasets:

$$Cosine\ Similarity\ (A,C) = \frac{1*1+1*1+1*1+1*0+0*1}{\|1*1+1*1+1*1+1*1\| \left\| \frac{1*1+1}{*1+1*1+1*1} \right\|} = 0.75$$

$$Cosine\ Similarity\ (A,E) = \frac{1*1+1*1+1*0+1*0+0*1+0*1}{\|1*1+1*1+1*1+1*1\| \atop \|1*1+1*1+1*1+1*1\|} = 0.50$$

$$TFCS\ (A,C) = \frac{10*10+5*5+6*6+6*0+0*4}{\|10*10+5*5+6*6+6*6\| \left\| \frac{10*}{10+5*5+6*6+4*4} \right\|} = 0.86$$

$$TFCS\ (A,E) = \frac{10*10+5*5+6*0+4*0+0*6+0*4}{\|10*10+5*5+6*6+6*6\| \atop \|10*10+5*5+4*4+4*4\|} = 0.71$$

Proof of TFCS as similarity metric: Similarity measure for any clustering algorithm should exhibit three main properties such as, (1) symmetry[d (i, j) = d(j, i)], (2) Non-negativity [s (i, j) ≥0] and (3) triangular inequality [s(i, j)≤s(i, k)+s(k, j)]. Term frequency based cosine similarity exhibits the following properties:

- Identity $TFCS(X,X) = 1$

$$CS(X,X) = \frac{\sum_{i=1}^{n} F_w(X_i) * F_w(X_i)}{\sqrt{\sum_{i=1}^{n}(F_w(X_i))^2} * \sqrt{\sum_{i=1}^{n}(F_w(X_i))^2}} = \frac{X}{X} = 1$$

- Symmetry: $TFCS(X,Y) = TFCS(Y,X)$

$$TFCS(X,Y) = \frac{\sum_{i=1}^{n} F_w(X_i)*F_w(Y_i)}{\sqrt{\sum_{i=1}^{n}(F_w(X_i))^2}*\sqrt{\sum_{i=1}^{n}(F_w(Y_i))^2}} = \frac{\sum_{i=1}^{n} F_w(Y_i)*F_w(X_i)}{\sqrt{\sum_{i=1}^{n}(F_w(Y_i))^2}*\sqrt{\sum_{i=1}^{n}(F_w(X_i))^2}} = TFCS(Y,X)$$

- Non Negativity: $TFCS(X,Y) \geq 0$

$R1: \forall_y Y \neq X \rightarrow TFCS(X,Y) < 1$
$(\theta))is\ the\ similarity$
$R2$: val(cos $measure\ between\ x\ and\ y, whose\ value\ lies\ between\ 0\ to\ 1$

$From\ R1 \cap R2 \rightarrow\ TFCS(X,Y) \geq 0$

$$TFCS(X,Y) = \frac{\sum_{i=1}^{n} F_w(X_i)*F_w(Y_i)}{\sqrt{\sum_{i=1}^{n}(F_w(X_i))^2}*\sqrt{\sum_{i=1}^{n}(F_w(Y_i))^2}} \geq 0$$

- **Triangular inequality:** The triangle inequality for cosine is same as for term frequency based cosine similarity. To rotate from x to y is to rotate to z and hence to y. The sum of those two rotations cannot be less than the rotation directly from x to y. $TFCS(X,Y) \leq TFCS(X,Z) + TFCS(Z,Y)$.

$$TFCS(X,Y) = \frac{\sum_{i=1}^{n} F_w(X_i)*F_w(Y_i)}{\sqrt{\sum_{i=1}^{n}(F_w(X_i))^2}*\sqrt{\sum_{i=1}^{n}(F_w(Y_i))^2}}$$

$$TFCS(X,Z) = \frac{\sum_{i=1}^{n} F_w(X_i)*F_w(Z_i)}{\sqrt{\sum_{i=1}^{n}(F_w(X_i))^2}*\sqrt{\sum_{i=1}^{n}(F_w(Z_i))^2}}$$

$$TFCS(Z,Y) = \frac{\sum_{i=1}^{n} F_w(Z_i)*F_w(Y_i)}{\sqrt{\sum_{i=1}^{n}(F_w(Z_i))^2}*\sqrt{\sum_{i=1}^{n}(F_w(Y_i))^2}}$$

$$TFCS(X,Z) + TFCS(Z,Y) = \frac{\sum_{i=1}^{n} F_w(X_i)*F_w(Z_i)}{\sqrt{\sum_{i=1}^{n}(F_w(X_i))^2}*\sqrt{\sum_{i=1}^{n}(F_w(Z_i))^2}} + \frac{\sum_{i=1}^{n} F_w(Z_i)*F_w(Y_i)}{\sqrt{\sum_{i=1}^{n}(F_w(Z_i))^2}*\sqrt{\sum_{i=1}^{n}(F_w(Y_i))^2}} = \frac{\sum_{i=1}^{n} F_w(X_i)*F_w(Y_i)*F_w(Z_i)}{\sqrt{\sum_{i=1}^{n}(F_w(X_i))^2}*\sqrt{\sum_{i=1}^{n}(F_w(Y_i))^2}*\sqrt{\sum_{i=1}^{n}(F_w(Z_i))^2}}$$

$$TFCS(X,Z) + TFCS(Z,Y) > TFCS(X,Y)$$

RESULTS AND DISCUSSION

Environment: Experiments were conducted on Intel core i5 processor with 2.4 GHz with 6GB DDR3 memory and 1000 GB HDD running Windows operating system. Program for cosine similarity and term frequency based cosine similarity computation written in java language.

Data set: Real life dataset, such as Congressional Vote is obtained from UCI machine learning repository (Lichman, 2013). Vote: Each tuple represent votes for each of the U.S. House of Representatives Congressmen. Number of instances is 435 and number of attributes is 17. It is classified into democrats (267) and republicans (168).

Cluster dendrogram: R software (R Development Core Team, 2013) used to plot the dendrogram representation of hierarchical clustering algorithm using the cosine similarity and term frequency based cosine similarity. In this study, we used only 100 instances of the vote data set for representing the dendrogram structure of the

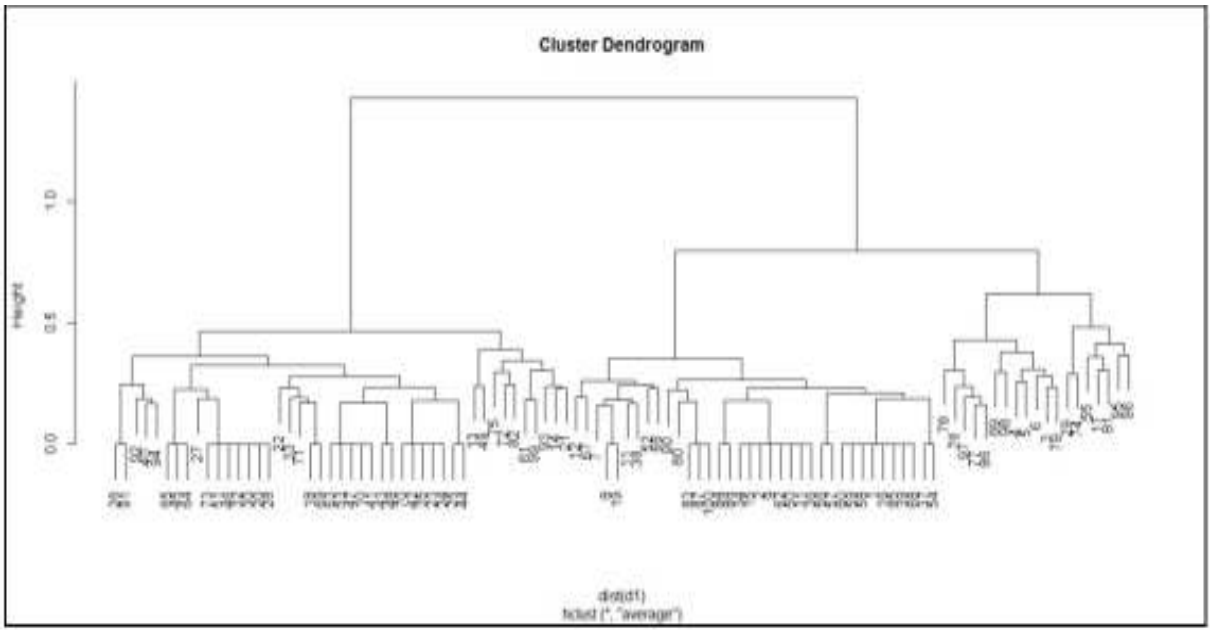

Fig. 1: Cluster dendrogram using TFCS

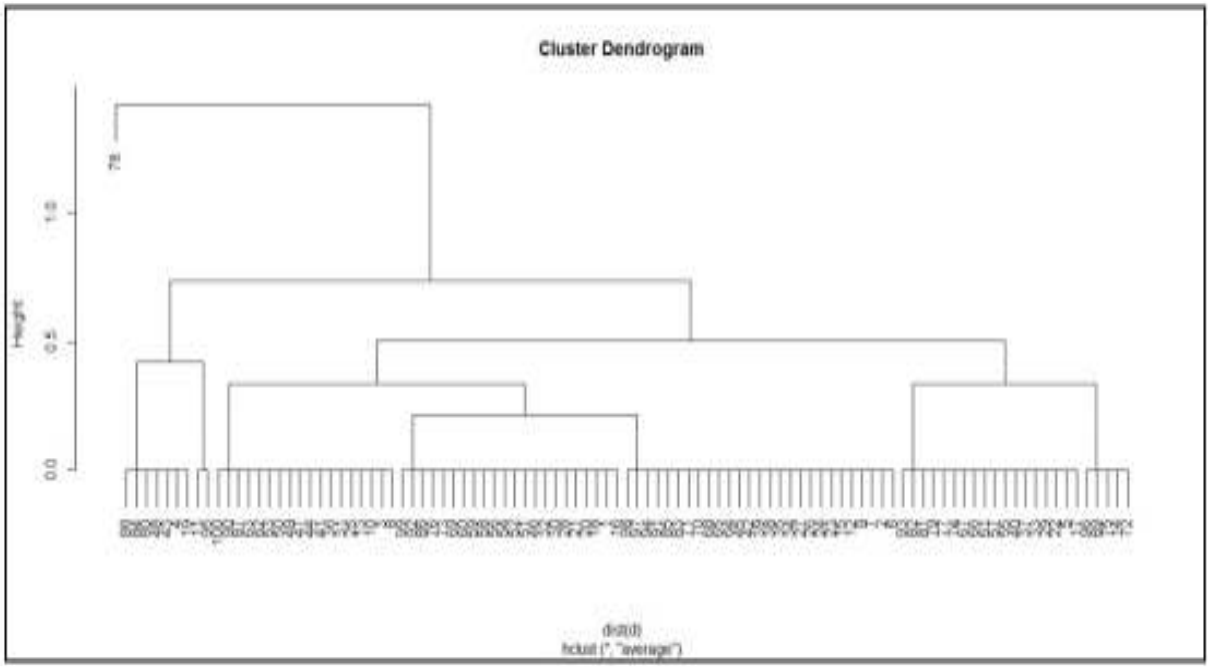

Fig. 2: Cluster dendrogram using cosine similarity

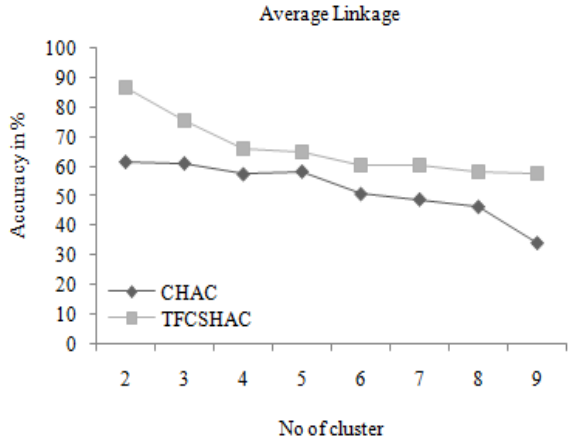

Fig. 3: Accuracy using average linkage

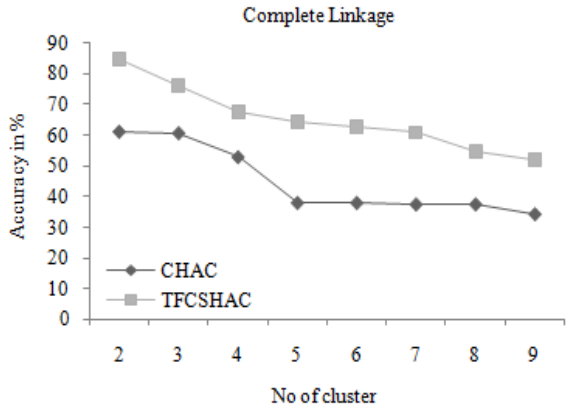

Fig. 4: Accuracy using complete linkage

hierarchical clustering algorithm. Figure 1 represents the dendrogram of TFCS based hierarchical algorithm and Fig. 2 represents the dendrogram of cosine based hierarchical algorithm.

Measure for cluster validation: The cluster validation is the process of evaluating the cluster results in a quantitative and objective manner. Cluster Accuracy 'r' is defined as:

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

where, 'n' refers number of instance in the dataset, 'ai' refers to number of instance occurring in both cluster i and its corresponding class and 'k' refers to final number of cluster. (2) Error rate 'E' is defined as:

$$E = 1 - r, \quad (5)$$

where, 'r' refers to the cluster accuracy.

Performance analysis on accuracy Vs no of cluster: Accuracy deals with how many instances are properly identified to the correct cluster. Experiments were conducted by varying the number of clusters from 2 to 9 using single, average and complete linkage. The graph plotted between accuracy and number of clusters is shown in Fig. 3 to 5. When cosine similarity measure is used with average linkage method of hierarchical clustering algorithm, accuracy of the cluster decreases sharply. The graph clearly shows the improvement for term frequency based cosine similarity in the sense that

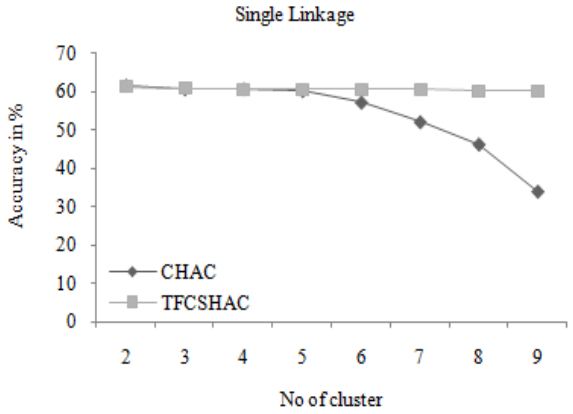

Fig. 5: Accuracy using single linkage

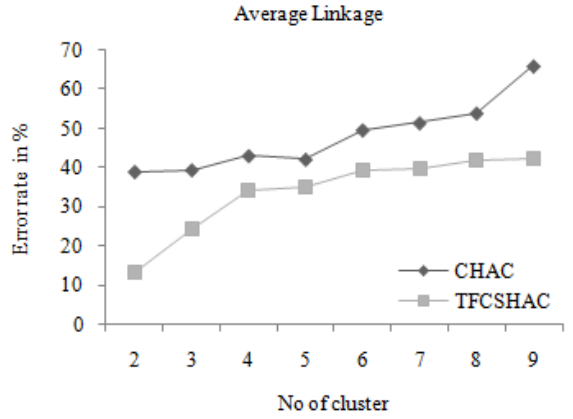

Fig. 6: Error rate using average linkage

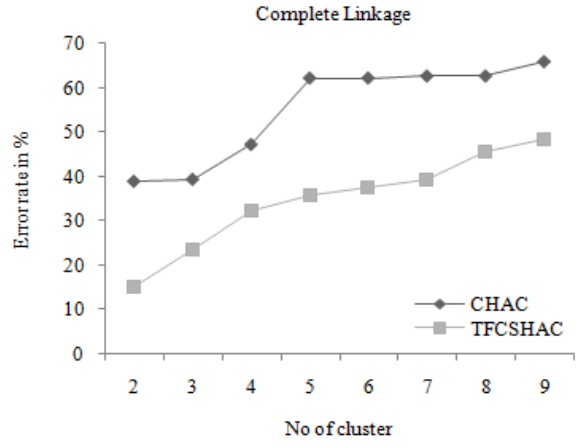

Fig. 7: Error rate using complete linkage

accuracy reaches a steady state if we increase the number of clusters above seven even though it drops slightly until the number of cluster is six.

The graph on Fig. 4 indicates, cosine similarity and term frequency based cosine similarity measure applied for hierarchical clustering algorithm using complete linkage. For all the iterations accuracy of TFCS based hierarchical clustering algorithm has higher accuracy than cosine similarity based hierarchical clustering algorithm.

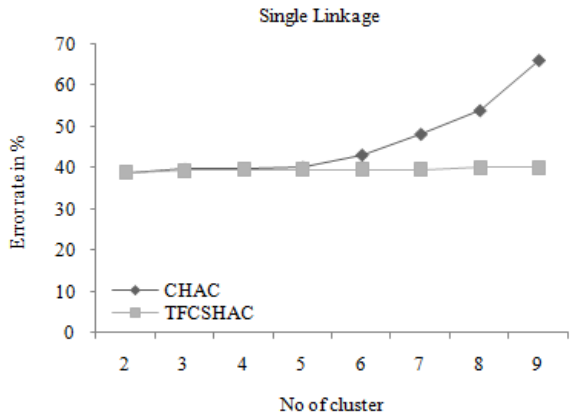

Fig. 8: Error rate using single linkage

The graph on Fig. 5 indicates, cosine similarity and term frequency based cosine similarity measure applied for hierarchical clustering algorithm using single linkage. For all the iterations, TFCS based hierarchical clustering algorithm shows a steady state of accuracy where as in cosine similarity based hierarchical clustering algorithm the accuracy drops down if the number of cluster is more than six. The average accuracy rate for cosine based hierarchical clustering algorithm using single linkage, complete linkage and average linkage is 59.94, 44.74 and 52.16%, respectively. The average accuracy rate for term frequency based cosine hierarchical clustering algorithm using single linkage, complete linkage and average linkage is 60.46, 65.29 and 66.32%, respectively.

The graph plotted between error rate and number of clusters is shown in Fig. 6 to 8. When cosine similarity measure is used with average linkage method of hierarchical clustering algorithm, error rate of the cluster increases sharply. The graph clearly shows for term frequency based cosine similarity, error rate reaches a steady state if we increase the number of clusters above seven.

The graph on Fig. 7 indicates, cosine similarity and term frequency based cosine similarity measure applied for hierarchical clustering algorithm using complete linkage. For all the iterations, cosine similarity of based hierarchical clustering algorithm has higher error rate than TFCS based hierarchical clustering algorithm.

The graph on Fig. 8 indicates, cosine similarity and term frequency based cosine similarity measure applied for hierarchical clustering algorithm using single linkage. For all the iterations, TFCS based hierarchical clustering algorithm shows a steady state of error rate where as in cosine similarity based hierarchical clustering algorithm the error rate increases if the number of cluster is more than six.

CONCLUSION

Categorical data are not analyzed as numerical data because of the absence of inherit ordering. The desirable property of cosine similarity is independent of document length. The performance of Term Frequency based Cosine Similarity measure for Hierarchical Clustering Algorithm (TFCSHCA) is evaluated against cosine based hierarchical clustering algorithm using categorical data. Results are evaluated for vote data set. Inferences from the proposed system are:

- In all the three cases (single, complete, average linkage method) term frequency based hierarchical clustering outperform the cosine based hierarchical clustering
- The Maximum accuracy attained for TFCS based single linkage, complete linkage and average linkage hierarchical clustering algorithm is 61, 85 and 87%, respectively.
- The Maximum accuracy attained for cosine similarity based single linkage, complete linkage and average linkage hierarchical clustering algorithm is 61, 60 and 61%, respectively.
- TFCS based average linkage hierarchical clustering algorithm achieves better result when compared with single or complete linkage methods.

REFERENCES

Agarwal, P., M.A. Alam and R. Biswas, 2010. Analysing the agglomerative hierarchical clustering algorithm for categorical attributes. Int. J. Innov. Manage. Technol., 1(2): 186-190.

Agresti, A., 1996. An Introduction to Categorical Data Analysis. Wiley, New York, Vol. 135.

Agresti, A., 2013. Categorical Data Analysis. John Wiley and Sons, New York.

Guha, S., R. Rastogi and K. Shim, 1998. CURE: An efficient clustering algorithm for large databases. SIGMOD Rec., 27(2): 73-84.

Guha, S., R. Rastogi and K. Shim, 1999. ROCK: A robust clustering algorithm for categorical attributes. Proceedings of the 15th International Conference on Data Engineering, pp: 512-521.

He, Z., X. Xu and S. Deng, 2002. Squeezer: An efficient algorithm for clustering categorical data. J. Comput. Sci. Technol., 17(5): 611-624.

Jiawei, H., K. Micheline and P. Jian, 2006. Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan Kaufmann, ISBN 0080475582, 9780080475585.

Karypis, G., E.H. Han and V. Kumar, 1999. Chameleon: Hierarchical clustering using dynamic modeling. Computer, 32(8): 68-75.

Lichman, M., 2013. UCI Machine Learning Repository. School of Information and Computer Science, University of California, Irvine, CA, Retrieved form: http://archive.ics.uci.edu/ml.

Lu, Y. and L.R. Liang, 2008. Hierarchical clustering of features on categorical data of biomedical applications. Proceeding of the ISCA 21st International Conference on Computer Applications in Industry and Engineering, pp: 26-31.

R Development Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved form: http://www.R-project.org/.

Santos, J.M., J.M. de Sá and L.A. Alexandre, 2008. Legclust-a clustering algorithm based on layered entropic subgraphs. IEEE T. Pattern Anal., 30(1): 62-75.

Virpioja, S., 2008. BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies. Retrieved from: http://www.cis.hut.fi/Opinnot/T-61.6020/2008/birch.pdf.