# Categorical Data Clustering using Frequency and Tf-Idf based Cosine Similarity

[1]S. Anitha Elavarasi, [2]J. Akilandeswari

[1]Department of Computer Science and Engineering, [2]Department of Information Technology
Sona College of Technology, Salem, Tamil Nadu, India

**Abstract-**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 due to the lack of inherit ordering. This paper describes Frequency and Tf-Idf Based Categorical Data Clustering (FTCDC) technique based on cosine similarity measure. The FTCDC system consists of four modules, such as data pre-processing, similarity matrix generation, cluster formation and validation. The System architecture of FTCDC is explained and its performance is examined using a simple example scenario. The performance on real world data is measured using accuracy and error rate. The performance of the system much relay on the similarity threshold selected for the clustering process.

## I. Introduction

Clustering is a process of grouping objects with similar properties [1]. Any clustering process should exhibit high intra class similarity and low inter class similarity. Data in real world are either numerical (continuous data with some ordering) or categorical (set of categories with no ordering) in nature. Categorical data are distributed in three ways: (1) binomial, (2) multinomial and (3) Poisson [3]. Applications of categorical data are: health care, education, biomedical, genetics and marketing fields.

Categorical data exhibit various similarity measures such as cosine similarity, Term frequency- Inverse document frequency (Tf-Idf), binary matching, overlap, hamming, edit distance and jaccard. Cosine similarity is a popular method for text mining. It is used for comparing the document and finds the closeness among the data points. If the similarity value is zero no similarity exist between the data element and if the similarity vale is 1 similarity exist between two elements. One desirable property of cosine similarity is, it is independent of document length. Term frequency (Tf) refers to the number of time a term occurs in the document and Inverse document frequency (IDF) revels the importance of the terms available in the corpus.

In this paper frequency and Tfidf based cosine similarity used as an objective measure for clustering categorical data. The system architecture of FTCDC (Frequency and Tf-Idf Based Categorical Data Clustering) algorithm and its performance is measured on real word dataset.

This paper is organized as follows. Section 2 discuses related work. Section 3 describes the proposed system architecture. Section 4 describes the experimental result and finally section 5 concludes the work.

## II. Related Work

K-means algorithm is a well-known partition clustering algorithm. It is efficient for processing larger data set, suitable for numerical data set and sensitive to outliers. The author extends the k means by using simple matching dissimilarity function suitable for categorical data [4]. Mode value is used instead of mean value and finally a frequency based method for updating the clustering process which reduces the cost function. K modes algorithm produce only local optima. K-prototype [5] algorithm combines both K-means and K-modes for mixed numerical and categorical data.

Downloaded from www.edlib.asdf.res.in

Squeezer is a clustering algorithm for categorical data [6]. The data structures involved are Cluster Summary and Cluster Structure. Summary holds pair of attribute value and their corresponding support. Cluster Structure (CS) holds the cluster and summary information. The Squeezer algorithm produces scalable and high quality cluster. It makes only one scan for the entire data set. The disadvantages of Squeezer algorithm is, quality of the cluster depends on the threshold value.

ROCK stands for Robust Clustering using links [7]. It is a Agglomerative hierarchy clustering. It uses links to measure similarity between data point. 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. Rock supports both Boolean and categorical data. Scalability of the algorithm depends on the sample size. QROCK [8] is a quicker version of ROCK algorithm. Two major variation of QROCK from ROCK are: (1) Final clusters are formed as the connected components of certain graph. (2) Similarity threshold is specified instead of desired number of cluster. The author in [9] made comparative evaluation of different similarity measure used for categorical data. This paper describes the performance of a variety of similarity measures in the context of a specific data mining task such as outlier detection. Fuzzy K-modes [10] is an extension to fuzzy K-means algorithm designed for categorical data. It generates fuzzy partition matrix which improves confidence degree of data in different cluster.

Data Intensive Similarity Measure for Categorical Data analysis (DISC) [11]. It makes use of a data structure called categorical information table (CI Table). Similarity matrix construction using DISC has 4 steps. (1)Initially similarity matrix is defined using overlap measure, (2) formation of CI table, (3) Computation of similarity value (4) Updating similarity matrix.

DILCA - DIstance Learning in Categorical Attribute is the measure used by the author [12,13]. 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.

## III. Proposed Work

### A. Proposed System Architecture

The system architecture for FTCDC (Frequency and Tf-Idf Based Categorical Data Clustering) is represented in Fig1. The FTCDC system consists of four modules, such as data preprocessing, similarity matrix generation, cluster formation and validation.

**Preprocessing:** Data in real world are incomplete, inconsistent and noisy. Quality of resultant cluster depends on the quality of input data given to the system. Data cleaning deals with filling missing values, handling noisy data and resolving inconsistencies in data. In this paper missing value are ignored and noise data are removed from the dataset.

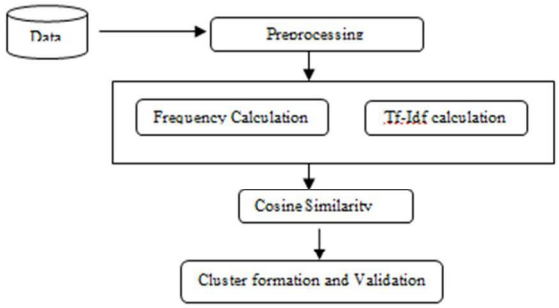


Figure1. System Architecture of FTCDC

Downloaded from www.edlib.asdf.res.in

**Similarity Computation**: It has four sub functions to be carried out for computing similarity. They are, (1) Frequency Computation: Frequency computation deals with calculating the frequency of occurrence for each attributes available in the dataset. (2) T Computation: It deals with calculating the Tf-Idf value for each attribute. (3) Occurrence Based Cosine: Occurrence information are generated and stored in a multi-dimensional array. Occurrence Based Cosine function deals with computing the similarity matrix using cosine similarity defined in (1). (4) T Cosine: It computes the similarity matrix using Tf-Idf based cosine similarity defined in (2).

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

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

**Cluster formation:** Cluster formation deals with grouping data with maximum similarity (greater the similarity closer the data points). Similarity matrix is constructed and threshold similarity values are selected. Select the data points whose similarity is less than the threshold and place them to a separate cluster.

**Cluster validation:** It is the process of evaluating the cluster results in a quantitative and objective manner. Generated clustered are validated using accuracy and error rate.

### B. Procedure

Input: Given Dataset with 'N' categorical attribute
Output: 'k' clusters
Algorithm: FTCDC
Begin
  Initialize n as total number of tuple
  Initialize occurrence of all attribute ($a_i$) and TF($a_i$) be zero
  read the data one by one
//Computer frequency and Tfidf
  while not EOF do
    for each attribute $a_i$ on A
      calculate the occurrence of each attribute as Ow ($a_i$)
      calculate the Tfidf for each attribute as TI ($a_i$)
    end for
  end while
//Similarity Matrix formation using FrequencyComputation and TComputation
  for i=1 to vectorlength do
    compute similarity Sim(X,Y) = XYvector/sqrt(Xvector) * sqrt(Yvector) using equation 1
  end for
  for i=1 to vectorlength do
    compute similarity Sim2(X,Y)= XYvector/sqrt(Xvector) * sqrt(Yvector) using equation 2
  end for
 //Cluster the data with max similarity using OccurranceBasedCosine and TCosine
   MaxSim = Select the maximum/ threshold similarity value
   Check for similarity matrix > MaxSim
     assign to the respective cluster Cluster[k++]
   return final cluster formed
end

Downloaded from www.edlib.asdf.res.in

### C. Sample Walkthrough

Assume the following sentences X,Y and Z. X : Rose is a beautiful flower, Y: Rose is red in color and Z: Rose is a flower. Similarity between X,Y and X,Z are calculated using equation 1 & 2 and results are shown in Table 1. Tf-Idf values act as an normalized weights for each of the attributes which can be applied to the normal cosine formula to get a better accuracy level.

Table I: similarity comparison

| Similarity between data point | Cosine Similarity | Occurrence based Cosine Similarity (OBCS) | TFIDF Similarity | Tf-Idf based Cosine Similarity |
|---|---|---|---|---|
| X,Y | 0.399 | 0.72 | 0.2 | 0.126 |
| X,Z | 0.894 | 0.965 | 0.51 | 0.568 |

## IV. Result and Discussion

Real life dataset Mushroom is obtained from UCI machine learning repository [14]. **Mushroom:** Each tuple represent the physical characteristic of mushroom. Number of instance is 8124 and number of attribute is 22.. In this paper FTCDC is validated using two of the external measure such as (1) Cluster Accuracy 'r' is defined a

$$r = \frac{\sum_{i=1}^{k} a_i}{n} \qquad (3)$$

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. (2) Error rate 'E' is defined as

E = 1- r, (4)

Where 'r' refers to the cluster accuracy. Occurrence based Cosine similarity outperforms other algorithms such as Rock, K modes and fuzzy K means. Tf-Idf based cosine similarity results are yet to compare with other algorithms.

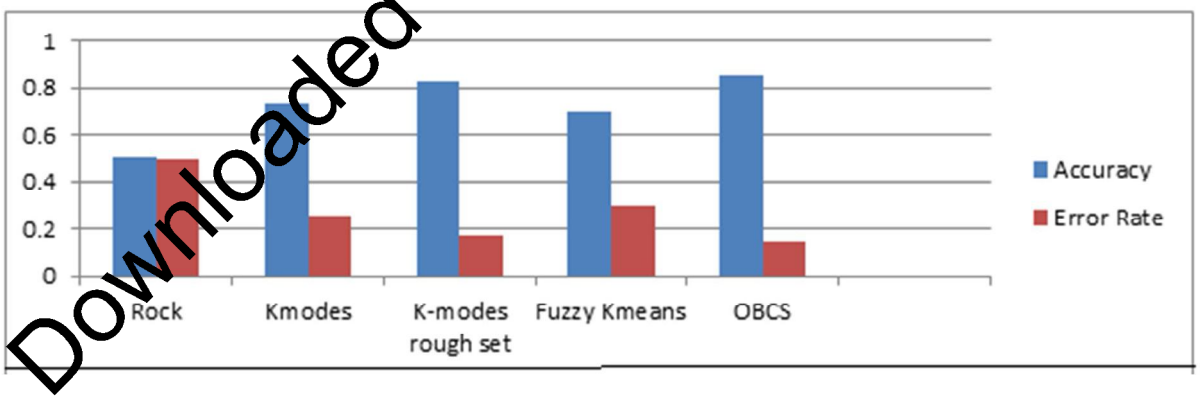


Figure 2.Accuracy and Error rate for Mushroom Data set

## V. Conclusion

This paper describes frequency and Tf-Idf based cosine similarity as an objective measure for categorical data clustering. The system architecture of FTCDC (Frequency of occurrence and Tf-Idf Based Categorical Data Clustering) algorithm is illustrated. Performance of OTCDC mainly depends on similarity value

Downloaded from www.edlib.asdf.res.in

(MaxSim). Cluster validation is performed using accuracy and error rate for mushroom dataset using different categorical clustering algorithm. The system can be further enhanced in future to achieve pure cluster formation and to reduce the execution time. Semantic information can be further added to the system to enhance its performance.

## References


[1] Jiawei Han, Micheline Kamber.; *Data Mining Concepts and Techniques.* The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, (2000).

[2] Hana Rezankova, *Cluster Analysis and categorical data statistika,* pp. 216-232, 2009.

[3] Alan Agresti, *An Introduction to categorical data analysis, second edition*, A John Wiley & sons publication

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

[5] Zhexue Huang; Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery 2,* pp283–304, 1998.

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

[7] 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.

[8] Dutta, M., A. Mahanta, and A. Pujari.; QROCK: A quick version of the ROCK algorithm for clustering of categorical data. Pattern Recognition Letters, vol. 26, pp2364-2373, 2005.

[9] Shyam Boriah, Varun Chandola, Vipin Kumar. Similarity Measures for Categorical Data: A Comparative Evaluation. In Proceedings of the eighth SIAM International Conference on Data Mining (2007).

[10] Ng, M.K. and Jing, L.; A new fuzzy k-modes clustering algorithm for categorical data. *Int. J. Granular Computing, Rough Sets and Intelligent Systems,* Vol. 1, No. 1, pp105–119, 2009

[11] Aditya Desai, Himanshu Singh, Vikram Pudi, DISC: Data Intensive Similarity for Categorical Attributes. *The 15thPacific-Asia Conference on Knowledge Discovery and Data Mining, Shenzen,* China, 2011.

[12] Dino Ienco, Ruggero G. Pensa And Rosa Meo.; From Context to Distance: Learning Dissimilarity for Categorical Data Clustering. *ACM Transactions on Knowledge Discovery from Data ,* 2011.

[13] D Ienco, R Pensa, R Meo.; Context-based distance learning for categorical data clustering. *Advances in Intelligent Data Analysis VIII*, pp. 83-94, 2009.

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


Downloaded from www.edlib.asdf.res.in