A Novel Cosine Similarity Like Data Clustering Method for Effective Data Classification in Data Mining

D. Mabuni

Abstract: In data mining ample techniques use distance based measures for data clustering. Improving clustering performance is the fundamental goal in cluster domain related tasks. Many techniques are available for clustering numerical data as well as categorical data. Clustering is an unsupervised learning technique and objects are grouped or clustered based on similarity among the objects. A new cluster similarity finding measure, which is cosine like cluster similarity measure (CLCSM), is proposed in this paper. The proposed cluster similarity measure is used for data classification. Extensive experiments are conducted by taking UCI machine learning datasets. The experimental results have shown that the proposed cosine like cluster similarity measure is superior to many of the existing cluster similarity measures for data classification.

Keywords: Clustering numerical data, clustering performance, cosine like cluster similarity, distance based measures.

I. INTRODUCTION

Clustering is a technique of grouping similar objects together based on their similarity features. Clustering is defined as the classification of data objects into homogeneous groups or clusters. Applications of clusters are: machine learning, data mining, science, biology, medicine, genomics, image analysis, pattern recognition, information retrieval, microbiology, defense, military. Distance functions are the most commonly used means for implementing cluster similarity measures. Distance based similarity measures are commonly used in distance-based clustering. Usage of similarity measures is easy in two or three dimensions but the complexity increases as the dimensionality increases. Definitely there is a need to find one standard framework for easy handling of high dimensional data. Similarity or distance measures are considered to be the core components frequently used by distance based clustering algorithms for clustering similar points into the same cluster and different points into the different clusters. Clustering is the most important technique for unsupervised learning. Similarity values are generally numeric values between 0 and 1. Here, 0 means no similarity and 1 means complete similarity. Many real time applications need some sort of similarity measures to find similarity between two objects. Rand index is a special index that is used for comparing accuracies of cluster similarity measures. Some cluster similarity measures are mostly recommended for high dimensional data and some other cluster similarity measures are recommended for low dimensional data only. In the analysis of data clustering literature a large number of techniques are available for classifying data objects based on data similarity or data dissimilarity measures. The aim of cluster analysis is to group or cluster the objects based on the features of the objects. Many cluster similarity finding measures are based on finding distances between objects. Some distance based similarity measures are:

1) Cosine similarity measure
2) Euclidean distance
3) Weighted Euclidean distance
4) Average distance
5) Manhattan distance
6) Minkowski distance
7) Chord distance
8) The Canberra distance metric
9) Triangle distance
10) Hamming distance
11) Jaccard similarity measure
12) Mahalanobis distance
13) Gaussian similarity measure

In the data mining literature many cluster similarity measures have been proposed by the researchers. Gaussian similarity based measures are also very famous for some applications. Euclidean distance measure is not preferable method for high dimensional data mining applications.

One can use cosine similarity to find similarity between two files. The cosine similarity is a numerical measure used for finding similarity between two objects. The objects may be anything like documents or user records or accounts and so on. The cosine cluster similarity measure is considered to be one of the best used similarity measures in the literature. Selecting the optimal cluster similarity measure depends on the particular data structure. In data clustering cosine similarity is a standard metric used for finding similarity between two objects. In general cosine similarity in cluster domain is a measure of similarity between two objects. In data mining this measure is used to find cohesion within clusters. Cosine similarity measure used for data clustering based on Euclidean distance which is one of the most widely used cluster similarity measure is generally preferable for clustering. Euclidean distance is not the best technique for handling objects in terms probabilities.

Revised Manuscript Received on May 23, 2020.

* Correspondence Author
D. Mabuni*, Department of Computer Science, Dravidian University, Kuppam, India. Email: mabuni.d@gmail.com

International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075, Volume-9 Issue-8, June 2020

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication
Advantages of cosine similarity measure are:

1) The most important feature in cosine cluster similarity measure is its low complexity
2) Cosine similarity measure is one of the most popular text similarity measures and it is predominantly used in document clustering.

II. RELATED WORK

Data management plays a key role in effective, correct and in time decision making particularly in business related activities. For efficient and effective data management, data clustering is a frequently used machine learning technique in many domains including data mining, big data analytics. Selecting the best cluster similarity measure is the success key for the success of any task including business related tasks. Brief introductions of some data similarity measures are given. They are:

A. Minkowski similarity measure

It is a collection of similarity measures including Euclidean distance and Manhattan distance. The minkowski distance is measured by the formula

$$d(min) = \left( \sum_{i=1}^{n} |x_i - y_i|^{m} \right)^\frac{1}{m}$$

Where $m$ is a positive real number and $x_i$ and $y_i$ are two vectors taken in n-dimensional space. This similarity finding method works well for both compact and isolated clustered datasets. The main disadvantage of this method is that large scale attributes generally dominates the small scale attributes and this problem is solved by normalizing the data values appropriately.

B. Manhattan Distance

Manhattan distance is a special case of minkowski distance when $m = 1$ and it is measured by using the formula

$$d(min) = \sum_{i=1}^{n} |x_i - y_i|$$

C. Euclidean Distance

Euclidean distance is a well-known distance measure used for similarity finding between objects. Minkowski distance becomes Euclidean distance when $m = 2$. It has all the drawbacks of minkowski distance measure.

D. Average Distance

Average distance is nothing but special case of modified version of Euclidean distance. In the case of n-dimensional space the formula is modified as

$$d(avg) = \left( \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i|^2 \right)^\frac{1}{2}$$

E. Weighted Euclidean Distance

When each attribute is given a special weight in the dataset then it is called weighted Euclidean distance and its formula is

$$d(weighted) = \left( \sum_{i=1}^{n} w_i |x_i - y_i|^2 \right)^\frac{1}{2}$$

F. Chord Distance

Chord Distance is another modified version of Euclidean distance. It overcomes all the drawbacks of Euclidean distance. Mahalanobis Distance, city block distance, and Pearson Correlation distance are also data similarity distance measures. Doaa S. Ali et. al. [1] proposed a new clustering method for clustering mixed datasets. This method works by selecting a specific similarity measure for each attribute. J. Kogan et. al. [2] proposed an optimization framework for generating k-means like data clustering algorithms including both batch k-means clustering algorithms and incremental clustering algorithms. J. A Irani et. al. [3] extensively studied clustering techniques and in their survey they have discussed all the cluster similarity measures in detail. L. Hamdad et. al. [4] proposed two similarity measures for spatial data clustering. L. Leydesdorff [5] compared Pearson coefficient with Salton’s cosine measure and experimentally concluded that cosine measure is insensitive to the number of zeros. Mohammad S, and Mohammadpour A, et. al. [6] proposed a new cluster similarity measure based on co variation coefficient and then evaluated performance of co variation similarity measure.

Reybob A et. al. [7] have proposed a new cluster data similarity measure for hierarchical clustering algorithms based on pitman measure of closeness. Pitman measure is a characteristic feature useful to find how much an estimator is close to its actual parameter. S. Sachdeva and B. Kastore [8] have clustered many English and Hindi datasets by using different types of data cluster similarity measures. After observing cluster results they concluded that cosine and Jaccard similarity measures are far better than many other clustering similarity measures. Sahar Sohanger and Dingding Wang, [9] thoroughly studied cluster similarity measures in particular cosine similarity measure and pointed out that this measure is not suitable for comparing similar objects in terms of their probabilities. They proposed a new sqrt-cosine similarity measure for finding similarity between two objects in areas such as document related tasks like clustering, queries, classification, association rule mining. Also large number of experiments is conducted to evaluate the performance measure of the proposed method. Document similarity is popularly using in many applications such as document clustering, query search, document classification, document summarization, fuzzy document clustering, fuzzy document classification and so on. Advantage of sqrt-cosine similarity measure is that it can handle high dimensional applications also. Shirkhorshidi AS et. al. [10] pointed out that many existing similarity finding measures are useful to handle 2 and 3 dimensions only and therefore there is a need of finding and standardizing high dimensional similarity finding measures. A technical and standard framework was proposed by the authors for comparing and evaluating performances of similarity measures based on distance based clustering algorithms. Abundant similarity measures are available for data
generalized framework consisting of categorical attribute
similarity measures. This framework includes five similarity
measures and authors have experimentally verified the
efficiency of the framework of cluster similarity measures.
Wen Zhang et. al. [12] proposed singular value
decomposition technique on clusters for improving the
discriminative power of latent semantic indexing.

III. PROBLEM DEFINITION

Finding the best cluster similarity measure for data
clustering is the main toughest problem in data mining.
Probably clustering is the most frequently used data mining
technique out of all the data mining techniques. No one cluster
similarity technique is always superior in all applications and
a particular cluster similarity technique is suitable for some
application and a separate cluster similarity technique is
needed in some other applications.

IV. PROPOSED COSINE LIKE CLUSTER
SIMILARITY MEASURE (CLCSM)

Cosine similarity measure is popularly used in many
applications for data clustering and data classification. A new
cluster similarity measure called CLCSM which is cosine like
cluster similarity measure is proposed and experimentally
employed in this paper. The proposed cluster similarity
technique is used for finding cluster similarity and then it is
used for data classification. The proposed cluster similarity
measure is computationally efficient and easy to implement
and produce accurate data classification results.

Cluster similarity measure is used for finding similarity
within and among the groupspor clusters formed based on
separate categorical value of each attribute in the dataset.
Suppose a particular attribute say “COURSE” has 4 distinct
categorical values then 4 sub groups or 4 clusters are created
and initially cluster similarity measure is computed within
each cluster separately and then finally similarity measure
among all these groups is measured. Finally attribute wise
aggregate similarity measures are considered for data
classification. Cosine like cluster similarity measure
(CLCSM) is computed using the equation (1)

\[
CLCSM = \frac{P + S}{\sqrt{P^2 + S^2 - 2 \cdot P \cdot S}}
\]  

(1)

Where P is the product and S is the sum of distinct classes of
each categorical value of the attribute A in the given training
dataset. For example, if the attribute COURSE has four distinct
categorical values, {B.Tech-CSE, B.Tech-ECE, B.Tech-EEE, Others}, then, 4 such cosine like cluster similarities are computed and then added to give up total similarity measurefor the selected attribute. The process is
repeated for each attribute in the given training dataset.

COURSE attribute has 4 sub clusters and CLCSM is
computed for each sub cluster separately and then finally all these 4 CLCSM sub scores are added for getting final total score of each attribute. That is

COURSE score = score-1 + score-2 + score-3 + score-4

Score-1, score-2, score-3, and score-4 are computed form sub
groups.

<table>
<thead>
<tr>
<th>Attribute value</th>
<th>1-class count</th>
<th>0-class count</th>
<th>CLCSM measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.Tech-CSE</td>
<td>8</td>
<td>4</td>
<td>2.2</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>8</td>
<td>4</td>
<td>2.2</td>
</tr>
<tr>
<td>B.Tech-EEE</td>
<td>8</td>
<td>4</td>
<td>2.2</td>
</tr>
<tr>
<td>Others</td>
<td>7</td>
<td>5</td>
<td>2.0435</td>
</tr>
<tr>
<td>Total score of COURSE attribute</td>
<td>8.6435</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CLCSM score for COURSE = “B.Tech-CSE” is computed
using the proposed cluster similarity measure shown in
equation (1)

\[
CLCSM = \frac{(8 \times 4) + (8 + 4)}{\sqrt{(8 \times 4)^2 + (8 + 4)^2 - 2 \cdot (8 \times 4) \cdot (8 + 4)}}
\]

Similarly

CLCSM score for COURSE = “B.Tech-ECE” = 2.2

Now CLCSM score for COURSE = “Others” is computed
using the same equation (1)

\[
CLCSM = \frac{44}{\sqrt{1024 + 144 - 768}}
\]

Now CLCSM scores for the attribute MLK are computed
using the equation (1)

<table>
<thead>
<tr>
<th>Attribute value</th>
<th>1-class count</th>
<th>0-class count</th>
<th>CLCSM measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>8</td>
<td>8</td>
<td>1.6666</td>
</tr>
<tr>
<td>Medium</td>
<td>15</td>
<td>1</td>
<td>31.0</td>
</tr>
<tr>
<td>High</td>
<td>8</td>
<td>8</td>
<td>1.6666</td>
</tr>
<tr>
<td>Total score of MLK attribute</td>
<td>34.3333</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CLCSM score for MLK = “High” is

\[
CLCSM = \frac{8 \times 8 + (8 + 8)}{\sqrt{(8 \times 8)^2 + (8 + 8)^2 - 2 \cdot (8 \times 8) \cdot (8 + 8)}}
\]
A Novel Cosine Similarity Like Data Clustering Method for Effective Data Classification in Data Mining

\[
CLCSM = \frac{64 + 16}{\sqrt{(64)^2 + (16)^2 - 2 \times 64 \times 16}}
\]

\[
CLCSM = \frac{80}{\sqrt{4096 + 256 - 2048}} = \frac{80}{\sqrt{2304}} = \frac{80}{48} = 1.6667
\]

CLCSM score for MLK = “Medium” is = 1.6667

CLCSM score for MLK = “Low” is

\[
CLCSM = \frac{(15 \times 1) + (15 + 1)}{\sqrt{(15 \times 1)^2 + (15 + 1)^2 - 2 \times (15 \times 1) \times (15 + 1)}}
\]

\[
CLCSM = \frac{31}{\sqrt{225 + 256 - 480}} = \frac{31}{1} = 31.0
\]

TABLE-3 TOS attribute measures using CLCSM

<table>
<thead>
<tr>
<th>Attribute value</th>
<th>1-class count</th>
<th>0-class count</th>
<th>CLCSM measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>15</td>
<td>9</td>
<td>1.4324</td>
</tr>
<tr>
<td>Yes</td>
<td>16</td>
<td>8</td>
<td>1.4615</td>
</tr>
</tbody>
</table>

Total score of TOS attribute = 2.894

CLCSM score for TOS = “No” is

\[
CLCSM = \frac{(15 \times 9) + (15 + 9)}{\sqrt{(15 \times 9)^2 + (15 + 9)^2 - 2 \times (15 \times 9) \times (15 + 9)}}
\]

\[
CLCSM = \frac{159}{\sqrt{18225 + 576 - 6480}} = \frac{159}{\sqrt{12321}}
\]

\[
CLCSM = \frac{159}{\sqrt{12321}} = \frac{159}{111} = 1.4324
\]

CLCSM score for TOS = “Yes” is

\[
CLCSM = \frac{(16 \times 8) + (16 + 8)}{\sqrt{(16 \times 8)^2 + (16 + 8)^2 - 2 \times (16 \times 8) \times (16 + 8)}}
\]

\[
CLCSM = \frac{152}{\sqrt{16384 + 576 - 6144}}
\]

CLCSM score for TOEFL = “No” is

\[
CLCSM = \frac{(24 \times 1) + (24 + 1)}{\sqrt{(24 \times 1)^2 + (24 + 1)^2 - 2 \times (24 \times 1) \times (24 + 0)}}
\]

\[
CLCSM = \frac{49}{\sqrt{576 + 625 - 1200}} = \frac{49}{49} = 49.0
\]

CLCSM score for TOEFL = “Yes” is

\[
CLCSM = \frac{(7 \times 17) + (7 + 17)}{\sqrt{(7 \times 17)^2 + (7 + 17)^2 - 2 \times (7 \times 17) \times (7 + 17)}}
\]

\[
CLCSM = \frac{(119) + (24)}{\sqrt{14161 + 576 - 5712}} = \frac{143}{\sqrt{9025}}
\]

\[
CLCSM = \frac{143}{\sqrt{9025}} = \frac{143}{95} = 1.5052
\]

TABLE-4 TOEFL attribute measures using CLCSM

<table>
<thead>
<tr>
<th>Attribute value</th>
<th>1-class count</th>
<th>0-class count</th>
<th>CLCSM measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>24</td>
<td>0</td>
<td>49.0</td>
</tr>
<tr>
<td>Yes</td>
<td>7</td>
<td>17</td>
<td>1.5053</td>
</tr>
</tbody>
</table>

Total score of TOEFL attribute = 50.5053

Maximum score value is generated for the TOEFL attribute. Hence, best attribute is TOEFL as a result of this data is classified based on distinct categorical values of TOEFL attribute. In a similar manner data classification is performed in the next iterations and the final resulted data classification.
model details are shown in FIGURE-1.

A. Datasets Description

Total 13 datasets are employed in this paper for experimental purpose. Out of 13 datasets 8 datasets are taken from standard UCI machine learning repository and 5 are manually created sample datasets.

1) Breast Cancer Dataset: It consists of 192 training instances and 95 testing instances. All these instances are described with 9 predictor attributes and one class label attribute. This dataset consists of only two class labels 0 and 1.

2) Nursery Dataset: It consists of 9719 training instances and 3240 testing instances. This set is described with 8 predictor attributes and 1 class attribute. There are five distinct class labels represented with 0, 1, 2, 3 and 4 respectively.

3) Car Evaluation: The training dataset size of the Car Evaluation is 1296 instances and dataset size of the testing dataset is 432 instances. The number of predictor attributes is 6 and there is one class label with 4 distinct classes denoted by 0, 1, 2 and 3 respectively.

4) Balance Scale: It consists of 469 training instances and 156 testing instances and it is described with 4 predictor attributes and one class label with 0, 1, and 2 classes. In the sample training dataset attribute descriptions are MLK (machine learning knowledge), TOS (test of scholarship), and TOEFL (test of English as foreign language).

TABLE-6 Stanford University Admission Dataset

<table>
<thead>
<tr>
<th>COURSE</th>
<th>MLK</th>
<th>TOS</th>
<th>TOEFL</th>
<th>ADMISSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.Tech-ECE</td>
<td>Low</td>
<td>Yes</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Low</td>
<td>No</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Low</td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>Yes</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>Yes</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>Yes</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>Yes</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>Yes</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>B.Tech-ECE</td>
<td>Medu</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
</tbody>
</table>

Class label, Admission = 1 means seat allotted and Admission = 0 means seat not allotted.

5) Primary Tumor: This dataset consists of training dataset of size 254 tuples and testing dataset of size 85 tuples. Number of predictor attributes is 17 and class label consists of 0, 1, and 2 values that represent types of tumors.

6) Lymphography: It consists of 112 instances of training dataset size and 36 instances of testing dataset size. There are 18 predictor attributes and one class attribute consists of class labels 1, 2, 3, 4, 5, 6, 7, and 8.

7) Hayes Roth: consists of 132 training tuples and 28 testing tuples. This dataset consists of 5 predictor attributes and one class attribute.

8) SPECT: The size of training dataset is 80 and the size of testing dataset is 187. This set consists of 22 predictor attributes and one class attribute. There are 2 classes represented with 0 and 1 labels.

All the remaining 5 datasets used in experimentation are created manually. All the training datasets contain only categorical attributes. A hypothetical dataset is employed in this paper for easy understanding purpose of proposed cluster similarity measure. Before creation of sample dataset 2 assumptions are taken into consideration during assigning class labels to each instance in the training dataset. University admission will be given to those students who qualified in TOEFL and having High or Medium machine learning knowledge.

V. ALGORITHM

Algorithm Cosine-Like-Cluster-Similarity (T, D, A)

Input:
T address of the first node
D set of training data instances
A set of attributes in the training dataset

Output:
Model of the data classifier
1. for i = 0 to (n-1) do
2. attribute-score[i] = 0
3. end for i
4. for each attribute \( A_i \) in the training dataset do
5. \( \text{score} = 0 \)
6. \( S = \text{find set of distinct categorical values of attribute } A_i \)
7. \( \text{sum} = 0, p = 0 \)
8. for each distinct categorical value \( C_j \) in \( S \) do
9. \( \text{sum} = \text{sum of class count of } C_j \)
10. \( p = \text{product of class counts of } C_j \)
11. \( \text{cosine-like-similarity} = \frac{(p+\text{sum})/(p*p+\text{sum}^2-2.0*p*\text{sum})}{\text{sum}} \)
12. \( \text{score} = \text{score} + \text{cosine-like-similarity} \)
13. end for j for loop
14. \( \text{attribute-score}[i] = \text{score} \)
15. end for i for loop
16. max = 0, location = 0
17. for each index \( i \) and the value in \( \text{attribute-score}[i] \) do
18. if(\( \text{attribute-score}[i] > \text{max} \)) then
19. \( \text{max} = \text{attribute-score}[i] \)
20. location = \( i \)
21. end if
22. end for index \( i \)
23. best-attribute = \( A[\text{location}] \)
24. return best-attribute

Algorithm Explanation:
Lines-1-3: score of each attribute is initialized to 0
Line-4: It is executed once for each attribute of the dataset
Line-5: It initializes score = 0. It adds sum of individual scores of each distinct categorical value of an attribute
Line-6: \( S \) is a set which stores distinct categorical values of each attribute \( A_i \)
Line-7: \( \text{sum} = \text{sum of class counts and } p = \text{product of class counts of distinct categorical value of each attribute} \)
Line-8-11: for each distinct categorical value \( C_j \) of the attribute \( A_i \), sum and product of class counts are computed then cosine like similarity is computed form sum and product.
Line-12: score is total sum of cosine like similarity values of each distinct categorical value of attribute \( A_i \)
Line-13: is the end of computations of each attribute \( A_i \)
Line-14: total score of each attribute \( A_i \) is stored separately in the array attribute-score[]
Line-15: end of score computations of all attributes
Line-16-22: for finding maximum score of all attributes of the dataset
Line-23: the attribute whose score is maximum, is selected and then returned to the calling function.

The proposed algorithm is executed on the given training dataset shown in TABLE-6 and the resulted data classification classifier is shown in the FIGURE-1. If the students are not qualified in TOEFL then they should be given admission in the university. That is, for getting university admission passing in TOEFL exam is prerequisite and the high or medium machine learning knowledge (MLK) is compulsory. For all students who have low machine learning knowledge will be given university admission only for non engineering students who have not qualified in the test of scholarship assistance (TOS) exam. Rules are created from the root to leaf paths. The rules generated from the output model are listed below:

Rule-1: If (TOEFL score < 60) then no admission
Rule-2: If (TOEFL score >= 60) and (MLK = high or medium) then admission will be given
Rule-3: If (TOEFL score >= 60) and (MLK = low) and (course = others) and (TOS = No) then admission will be given
Rule-4: If (TOEFL score >= 60) and (MLK = low) and (course = others) and (TOS = Yes) then no admission

VI. EXPERIMENTS

Experiments are conducted by taking 13 datasets of which 9 datasets are taken from UCI machine learning repository and the remaining 4 datasets are imaginary datasets created manually in this paper. All the 13 datasets are experimented by running the proposed algorithm and the results are tabulated in the TABLE-7. From the tabulated results it is clear that the proposed algorithm has produced far better results than the best existing C4.5 algorithm in all the cases except in the case of CAR evaluation experiment where C4.5 algorithm has produced the best result with an accuracy of 81.25.

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Training Data Size</th>
<th>Test Data Size</th>
<th>C4.5</th>
<th>CLCSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>BreastCancer</td>
<td>192</td>
<td>95</td>
<td>61.052</td>
<td>80.0</td>
</tr>
<tr>
<td>Lymphography</td>
<td>112</td>
<td>36</td>
<td>27.777</td>
<td>44.44</td>
</tr>
<tr>
<td>Primary Tumor</td>
<td>254</td>
<td>85</td>
<td>55.294</td>
<td>71.764</td>
</tr>
<tr>
<td>Hayes-Roth</td>
<td>132</td>
<td>28</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>SPECT</td>
<td>80</td>
<td>187</td>
<td>58.823</td>
<td>60.96</td>
</tr>
<tr>
<td>Balance scale</td>
<td>469</td>
<td>156</td>
<td>45.512</td>
<td>66.026</td>
</tr>
<tr>
<td>NurseryData</td>
<td>9719</td>
<td>3240</td>
<td>86.358</td>
<td>92.932</td>
</tr>
<tr>
<td>CAR</td>
<td>1296</td>
<td>432</td>
<td>81.25</td>
<td>80.555</td>
</tr>
<tr>
<td>All Electronics</td>
<td>14</td>
<td>14</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>BP_SUGAR</td>
<td>36</td>
<td>36</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Job Dataset</td>
<td>90</td>
<td>90</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Loan Dataset</td>
<td>80</td>
<td>40</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

FIGURE-1 Data Classification Model
University Admission  |  48 | 48 | 100 | 100

FIGURE-2 Classification Accuracies of datasets
Classification accuracy of CLCSM is greater than C4.5 in all the cases except CAR data. For small sets accuracy 100 percent but for larger sets accuracy is less than 100 percent.

VII. CONCLUSION
The main challenge facing by many researchers and professionals today is that how to select the correct cluster similarity measure or distance measure for effective data clustering. The main goal of data clustering is to find clusters or groups that are both homogeneous and well separated units. Different cluster similarity measures vary as the dimensionality of the dataset increases. There is a need to find more generalized framework of cluster similarity measures. Clustering is a popular data mining technique. There exists variety of similarity based measures for efficient and effective clustering. A new technique is proposed for data classification based on clustering that uses cosine like similarity approach. There is a broad scope for enhancing many of the existing cluster similarity measures. In the future efficient and effective similarity based methods with the state of the art features will be investigated for producing excellent outperforming data clustering and data classification results.

REFERENCES
12. Wen Zhang, Fan Xiao, Bin Li,and Siguang Zhang, Using SVD on Clusters to Improve Precision of Interdocument Similarity Measure”, computational intelligence and neuroscience, research article, volume 2016, ID 1096271, 11 pages.

AUTHORS PROFILE

D. Mabuni, completed M.Sc. (Computer Science), MCA and M.Phil. (Computer Science). Currently working as Assistant Professor in the Department of Computer Science at Dravidian University, Kuppan, Andhra Pradesh, India. My interested research areas are Data Mining, Databases, and User Interfaces.