MEASUREMENT OF SIMILARITY USING LINK BASED CLUSTER APPROACH FOR CATEGORICAL DATA

A PROJECT REPORT

Submitted by

PAVITHRA M

in partial fulfillment for the requirement of award of the degree of

MASTER OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

Department of Computer Science and Engineering
KUMARAGURU COLLEGE OF TECHNOLOGY,
COIMBATORE 641 049
(An Autonomous Institution Affiliated to Anna University, Chennai)

APRIL 2013

BONAFIDE CERTIFICATE

Certified that this project work titled **"MEASUREMENT OF SIMILARITY USING LINK BASED CLUSTER APPROACH FOR CATEGORICAL DATA"** is the bonafide work of Mrs. PAVITHRA M who carried out the research under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other students.

Prof.N.JAYAPATHI M.E.,	Dr.D.CHANDRAKALA M.E.,Ph.D.,
HEAD OF THE DEPARTMENT	**SUPERVISOR**
Dept of Computer Science and Engineering	**Associate Professor**
Kumaraguru College of Technology	Dept of Computer Science and Engineering
Coimbatore- 641 049	Kumaraguru College of Technology
	Coimbatore- 641 049

Submitted for the Project Viva-Voce examination held on _____.

----------	----------
Internal Examiner	**External Examiner**

ABSTRACT

Clustering is to categorize data into groups or clusters such that, the data in the same cluster are more similar to each other than to those in different clusters. The problem of clustering categorical data is to find a new partition in dataset. The underlying ensemble-information matrix presents only cluster-data point relations, with many entries being left unknown. This problem degrades the quality of the clustering result. A new link-based approach, which improves the conventional matrix by discovering unknown entries through similarity between clusters in an ensemble and an efficient link-based algorithm is proposed for the underlying similarity assessment. C-Rank link-based algorithm is used to improve clustering quality and ranking clusters in weighted networks. C-Rank consists of three major phases: (1) identification of candidate clusters; (2) ranking the candidates by integrated cohesion; and (3) elimination of non-maximal clusters. Finally apply this clustering result in graph partitioning technique is applied to a weighted bipartite graph that is formulated from the refined matrix.

ஆய்வுச்சுருக்கம்

தொகுப்பு குழுக்கள் அல்லது அதே கொத்து தரவு வெவ்வேறு கொத்தாக ஒருவருக்கொருவர் போன்ற கொத்தாக தரவை வகைப்படுத்தும் வகையில் உள்ளது. ஆணித்தரமான தரவு கொத்தாக்கலும் பிரச்சினையை சமாளிக்க தரவு ஒரு புதிய பகிர்வு கண்டுபிடிக்க வேண்டும், இதன் விளைவாக இந்த உத்திகள் முழுமையில்லாத தகவலை அடிப்படையாக கொண்டு ஒரு இருதி தரவு பகிர்வு உருவாக்க கவனிக்கப்பட்டது. அடிப்படை குழும தகவல் அணி பல உள்ளீடுகளை கொண்டு, ஒரே கொத்து தரவு புள்ளி உறவுகளை அளிக்கிறது, இந்த சிக்கலை தொகுப்பு விளைவின் தரத்தை குறைக்கிறது. வழக்கமான அணி ஒரு குழும உள்ள கொத்தாக மற்றும் ஒரு திறமையான இணைப்பு சார்ந்த படிமுறை இடையே ஒற்றுமை மூலம் தெரியவில்லை. உள்ளீடுகளை கண்டுபிடிப்பதன் மூலம் தொகுப்பு தர ஒரு புதிய இணைப்பு அடிப்படையிலான அணுகுமுறையை மேம்படுத்த அடிப்படை ஒற்றுமை மதிப்பீடு முன்மொழியப்பட்டது. இந்த காகித இணைப்பு சார்ந்த படிமுறை நிறை வலையமைப்பு தொகுப்பு தரம் மற்றும் தரவரிசை கொத்தாக மேம்படுத்த தரவரிசை முன்மொழிய தரவரிசை மூன்று முக்கிய கட்டங்கள் உள்ளன:வேட்பாளர் கொத்துகள் (1)அடையாளம் (2) ஒருங்கிணைந்த ஒட்டுத்தன்மைக்கான மூலம் வேட்பாளர்கள் இடத்தையும் மற்றும் அல்லாத அதிகபட்ச கொத்துகள் (3)நீக்கப்படும்.

ACKNOWLEDGEMENT

First and foremost, I would like to thank the Lord Almighty for enabling me to complete this project. I express my profound gratitude to **Padmabhusan Arutselvar Dr.N.Mahalingam, B.Sc., F.I.E , Chairman**, **Dr.B.K. Krishnaraj Vanavarayar**, **Co-Chairman, Mr. M. Balasubramaniam, M.Com., M.B.A, Correspondent, Mr.Sankar Vanavarayar**, **M.B.A., PGDIEM, Joint Correspondent** and **Dr.S.Ramachandran Ph.D., Principal** for providing the necessary facilities to complete my project.

I take this opportunity to thank **Prof.N.Jayapathi M.Tech**., Head of the Department, Department of Computer Science and Engineering, for his support and motivation. Special thanks to my Project Coordinator **Dr.V.Vanitha M.E., Ph.D.**, Senior Associate Professor, Department of Computer Science and Engineering, for arranging brain storming project review sessions.

I register my sincere thanks to my Guide **Dr.D.Chandrakala M.E., Ph.D.,** Associate Professor, Department of Computer Science and Engineering. I am grateful for her support, encouragement and ideas. I would like to convey my honest thanks to all **Teaching** and **Non Teaching Staff** members of the department and my classmates for their support.

I dedicate this project work to my **Parents** for no reasons but feeling from bottom of my heart that without their love, this work would not be possible.

- PAVITHRA M

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CO+SL	-	Co association with Single Link
CO+AL	-	Co association with Average Link
CSPA	-	Cluster based Similarity Partitioning Algorithm
EM	-	Expectation Maximization
HBGF	-	Hybrid Bipartite Graph Formulation
HGPA	-	Hyper Graph Partitioning Algorithm
IBGF	-	Instance Based Graph Formulation
KDD	-	Knowledge Discovery in Databases
LAC	-	Locally Adaptive Clustering
MCLA	-	Meta Clustering Algorithm
NMI	-	Normalized Mutual Information
RM	-	Refined Matrix
WTQ	-	Weighted Triple Quality

CHAPTER 1

INTRODUCTION

1.1. DATA MINING

Data mining is an integral part of knowledge discovery in databases (KDD), which is the overall process of converting raw data into useful information from large repositories.

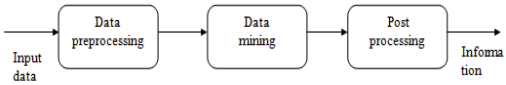


Fig.1.1 KDD process

Data mining tasks usually grouped into two types: dividing objects into groups (clustering i.e. descriptive) and assigning particular objects to these groups (classification i.e. predictive). Data mining, a synonym to "knowledge discovery in databases" is a process of analyzing data from different perspectives and summarizing it into useful information. It is a process that allows users to understand the substance of relationships between data which is shown in Fig.1.1.

Three important components of data mining systems are databases, data mining engine and pattern evaluation modules.

DATA CLUSTERING

Clustering is a process of partitioning a set of data (or objects) in a set of meaningful sub-classes, called clusters. Clustering helps users to understand the natural grouping or structure in a data set. Clustering is an unsupervised classification and it has no predefined classes. The requirements of clustering methods are,

- Scalability
- Dealing with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- The curse of dimensionality
- Interpretability and usability

Issues with clustering algorithms

There are two significant challenges inherent to clustering algorithms. First, various clustering algorithms find different structures (e.g., size, shape) in the same dataset. This is because each individual clustering algorithm has its own preferences due to the optimization of different criteria. Second, a single algorithm with different parameter settings can find various structures on the same dataset (Boulis,

2004). Since no labelled data are available, no cross-validation can be used to tune the parameters.

1.2.1 CLUSTERING METHODOLOGIES

- Connectivity based clustering (hierarchical clustering)
- Centroid-based clustering
- Distribution-based clustering
- Density-based clustering

1.2.1.1 Connectivity based clustering (hierarchical clustering)

Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. As such, these algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters do not mix which is shown in the Fig.1.2. Connectivity based clustering is a whole family of methods that differ by the way distances are computed (Boulis, 2004).

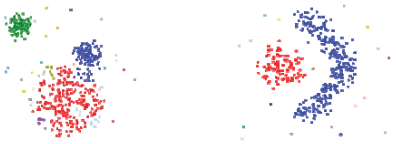


Fig.1.2. Linkage clustering examples

1.2.1.2 Centroid-based clustering

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set (Gibson, 2000). When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the k cluster centres and assign the objects to the nearest cluster centre, such that the squared distances from the cluster are minimized which is shown in Fig.1.3.

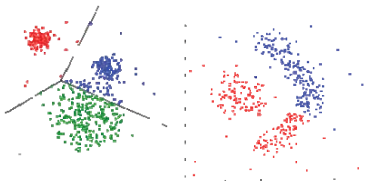


Fig.1.3 k-means clustering examples

1.2.1.3 Distribution-based clustering

The clustering model most closely related to statistics is based on distribution models which is shown in Fig.1.4. Clusters can then easily be defined as objects belonging most likely to the same distribution(Gibson, 2000). Distribution-based clustering is a semantically strong method, as it not only provides you with clusters, but also produces complex models for the clusters that can also capture correlation and dependence of attributes.

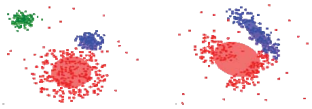


Fig.1.4 EM clustering examples

1.2.1.4 Density-based clustering

In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points which is shown in Fig.1.5. A cluster consists of all density-connected objects (which can form a cluster of an arbitrary shape, in contrast to many other methods) plus all objects that are within these objects' range (A.P. Topchy, 2005).

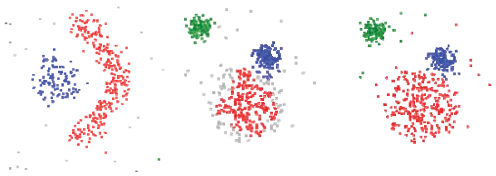


Fig.1.5 Density-based clustering examples

Clustering has wide applications in Pattern Recognition, Spatial Data Analysis, Image Processing, Market Research, Information Retrieval, Web mining, Marketing, Biology.

1.3 The Curse of Dimensionality

The curse of dimensionality refers to the increase in the sparsity of data as dimensionality increases (A.P.Topchy, 2005). In high dimensional spaces, finding regions of dense points becomes a difficult task. In low dimensional space, clusters can be found easily and patterns can be easily recognized. Data in only one dimension is relatively packed. In order to avoid the high dimensionality problem, dimensionality reduction techniques such as feature transformation and feature selection are used which is shown in Fig.1.6.

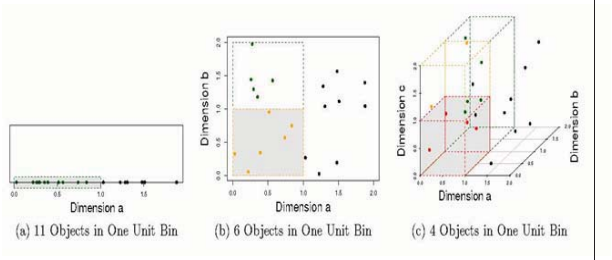


Fig.1.6 The Curse of Dimensionality

1.3.1 **Subspace Clustering**

Subspace clustering seeks to find clusters in a dataset by selecting the most relevant dimensions for each cluster separately. Subspace clustering methods searches various subspaces to find clusters and is well suited for high dimensional spaces which is shown in Fig.1.7. The two major types of search algorithms based on density:

- Top-down search

- Bottom-up search

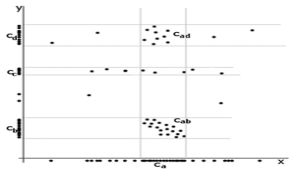


Fig.1.7 Subspace Clustering

1.4 LITERATURE REVIEW

1.4.1. Temporal Data Clustering via Weighted Clustering Ensemble With Different Representations

Yun Yang (2011) discuss an introduction to temporal data clustering. Temporal data clustering provides keystone techniques for discovering the intrinsic structure and condensing information over temporal data.

1.4.1.1. Temporal data representation:

Temporal data representations are generally classified into two categories:

- Piecewise representations.
- Global representations.

1.4.1.1.1. Piecewise Representation:

A piecewise representation is generated by partitioning the temporal data into segments at critical points based on a criterion, and then, each segment will be modelled into a concise representation.

1.4.1.2. Weighted Consensus Function:

The basic idea of weighted consensus function is the use of pair wise similarity between objects in a partition for evident accumulation, where a pair wise

similarity matrix is derived from weighted partitions and weights are determined by measuring the clustering quality with different clustering validation criteria.

1.4.1.3. Normalized Mutual Information (NMI):

The NMI is proposed to measure the consistency between two partitions, i.e., the amount of information (common structured objects) shared between two partitions.

Disadvantage

- ✓ It cannot achieve model selection.
- ✓ Grouping process cannot be achieved.

1.4.2 Cluster Ensembles for High Dimensional Clustering: An Empirical Study

Xiaoli Z (2005) discuss an introduction to cluster ensembles for high dimensional data clustering. It examine three different approaches to constructing cluster ensembles. To address high dimensionality, focus on ensemble construction methods that build on two popular dimension reduction techniques, random projection and principal component analysis (PCA).

1.4.2.1. Hybrid Bipartite Graph Formulation:

The graph edges can only connect instance vertices to cluster vertices, resulting a bipartite graph. But generally it is computationally more expensive than the other two graph formulation approaches.

1.4.2.1.1. Description of HBGF:

Given a cluster ensemble $\pi = \{\pi_1 \ldots\ldots, \pi_R\}$, HBGF constructs a graph G = (V,W), where

• $V = V_C \cup V_I$, where V_C contains t vertices each representing a cluster of the ensemble, V_I contains n number of vertices each representing an instance of the data set X.

• W is defined as the vertices i and j are both clusters or both instances, W(i, j) = 0; otherwise if instance i belongs to cluster j, W(i, j) = W(j, i) = 1 and 0 otherwise.

1.4.2.2. Instance-Based Graph Formulation(IBGF):

The Instance-Based Graph Formulation (IBGF) approach constructs a graph that models the pair wise relationship among instances of the data set X. IBGF constructs a fully connected graph with n edges, where n is the number of instances.

1.4.2.3. Cluster-Based Graph Formulation (CBGF):

Cluster-Based Graph Formulation (CBGF) constructs a graph that models the correspondence (similarity) relationship among different clusters in a given ensemble and partitions the graph into groups such that the clusters of the same group correspond(are similar) to one another.

Disadvantage

- ✓ Multiple low dimensional representation of data cannot be achieved.
- ✓ Fail to provide satisfactory performance.

1.4.3 A Framework for Cluster Ensemble Based on a Max Metric as Cluster Evaluator

Hosein Alizadeh(2003) discuss about a new criterion for clusters validation is proposed and it is based on the new cluster validation criterion a clustering ensemble framework is proposed. The main idea behind the framework is to extract the most stable clusters in terms of the defined criteria.

1.4.3.1. Max Method:

A drawback of computing stability is introduced and an alternative approach is suggested which is named Max method. It shows two primary partitions for which the stability of each cluster is evaluated.

Disadvantage

- ✓ Subset of selecting primary cluster or partitions cannot be achieved.
- ✓ Data clustering or unsupervised learning is very difficult problem.

1.4.4 Weighted Clustering Ensembles

Muna Al-Razgan(2007) discuss an introduction to weighted clustering. Cluster ensembles offer a solution to challenges inherent to clustering arising from its ill-posed nature. Cluster ensembles can provide robust and stable solutions by

leveraging the consensus across multiple clustering results, while averaging out emergent spurious structures that arise due to the various biases to which each participating algorithm is tuned.

1.4.4.1. Locally Adaptive Clustering

Let us consider a set of *n* points in some space of dimensionality D. A weighted cluster C is a subset of data points, together with a vector of weights $w = (w_1, \ldots, w_D)t$, such that the points in C are closely clustered according to the L_2 norm distance weighted using w.

1.4.4.1.1. Definition of Locally Adaptive Clustering

Given a set S of n points $x \in R_d$, a set of k centres $\{c_1, \ldots, c_k\}$, $c_j \in R_d$, $j = 1, \ldots, k$, coupled with a set of corresponding weight vectors $\{w_1, \ldots, w_k\}$, $wj \in R_d$, $j = 1, \ldots, k$, partition S into k sets.

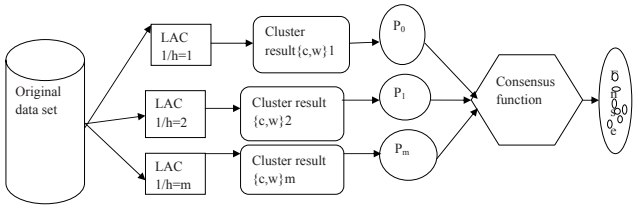


Fig.1.8 The clustering ensemble process

Disadvantage

- ✓ Difficult to combining multiple weighted clusters which belong to different subspace.
- ✓ Difficult to find k in an automated fashion through a cluster ensemble.
- ✓ Difficult to estimate the weighted vector w for each cluster in the dataset.

1.4.5 Combining Multiple Clustering System

Constantinos Boulis(2002) discuss an introduction to multiple clustering. Three methods for combining multiple clustering systems are presented and evaluated, focusing on the problem of ending the correspondence between clusters of different systems.

1.4.5.1. Correspondence problem

Each system is represented by a D*D matrix (D is the total no of observations) where the (i; j) position is either 1 if observations i and j belong to the same cluster and 0 otherwise. The average of all matrices is used as the input to a final similarity-based clustering algorithm. It has quadratic memory and computational requirements.

1.4.5.2. Hyper graph cutting problem

Each one of the clusters of each system is assumed to be a hyper edge in a hyper graph. The problem of finding consensus among systems is formulated as partitioning a hyper graph by cutting a minimum number of hyper edges.

Disadvantage

- ✓ Difficult to find correspondence between clusters of different systems.
- ✓ Optimization problem.

1.4.6 Clustering the Mixed Numerical and Categorical Datasets Using Similarity Weight and Filter Method

Srinivasulu Asadi(2006) discuss an introduction to filter method. Clustering is a challenging task in data mining technique. The aim of clustering is to group the similar data into number of clusters. Various clustering algorithms have been developed to group data into clusters. However, these clustering algorithms work effectively either on pure numeric data or on pure categorical data, most of them perform poorly on mixed categorical and numerical data types in previous k-means algorithm was used but it is not accurate for large datasets.

1.4.6.1. Filter Algorithm:

The original dataset is divided into two sub-datasets i.e., pure categorical dataset and the pure numerical dataset.

1.4.6.2. Steps of Filter Algorithm:

Step 1: Start with a tree built by the sequential initialization.
Step 2: Calculate mean and standard deviation of the edge weights distance array.
Step 3: Use their sum as the threshold.

Step 4: Perform multiple runs of Similarity Algorithm.
Step 5: Identify longest edge using Similarity.
Step 6: Remove this longest edge.
Step 7: Check Terminating Condition and continue.
Step 8: Put that number of clusters into Filter Method.

Disadvantage

- ✓ Efficient partitioning of a large data set into homogeneous groups or clusters cannot be achieved.
- ✓ Effective interpretation of clusters cannot be achieved.

CHAPTER 2

IMPLEMENTATION

2.1 SYSTEM ANALYSIS

2.1.1 Existing System

Many categorical data clustering algorithms have been introduced in recent years, with applications. The initial method was developed by making use of Gower's similarity coefficient the k-modes algorithm is proposed to extended the conventional k-means with a simple matching dissimilarity measure and a frequency-based method to update centroids (i.e., clusters' representative). As a single-pass algorithm, makes use of a pre specified similarity threshold to determine which of the existing clusters (or a new cluster) to which a data point. LIMBO is a hierarchical clustering algorithm that uses the Information Bottleneck (IB) framework to define a distance measure for categorical tuples. The concepts of evolutionary computing and genetic algorithm have also been adopted by a partitioning method for categorical data. Different graph models have also introduced by the STIRR, ROCK, and CLICK techniques. To resolve clustering categorical data partition problem different algorithms were introduced but still

there is problem. Cluster ensembles have emerged as an effective solution that is able to overcome these limitations, and improve the robustness as well as the quality of clustering results.

Drawbacks

- ➢ Minimizes error measures.
- ➢ Improving the accuracy.

2.1.2 Proposed System:

A new link-based approach, which improves the conventional matrix by discovering unknown entries through similarity between clusters in an ensemble and an efficient link-based algorithm is proposed for the underlying similarity assessment. To extend the work by analyzing the behavior of other link-based similarity measures with this problem the quality of the clustering result. C-Rank link-based algorithm is used to improve clustering quality and ranking clusters in weighted networks. C-Rank consists of three major phases:

- ➢ Identification of candidate clusters;
- ➢ Ranking the candidates by integrated cohesion; and
- ➢ Elimination of non-maximal clusters.

The clustering result in graph partitioning technique is applied to a weighted bipartite graph that is formulated from the refined matrix. To capture strength of weighted clusters is tricky, since edge weights have to be taken into account as well.

2.2 Problem Definition :

Cluster ensembles methods introduced to solve clustering problem

- The feature-based approach that transforms the problem of cluster ensembles to clustering categorical data (i.e., cluster labels).
- The direct approach that finds the final partition through relabeling the base clustering results.
- Graph-based algorithms that employ a graph partitioning method
- T he pair wise-similarity approach that makes use of co-occurrence relations between data points.

These methods generate the final data partition based on incomplete information of a cluster ensemble. As a result, the performance of existing cluster ensemble techniques may consequently be degraded as many matrix entries are left unknown.

2.3 Overview of The Project :

Clustering is a problem of great practical importance that has been the focus of substantial research in several domains for decades. It is defined as the problem of partitioning data objects into groups, such that objects in the same group are similar, while objects in different groups are dissimilar. This definition assumes that there is some well defined notion of similarity, or distance, between data objects. When the objects are defined by a set of numerical attributes, there are natural definition of distance based on geometric analogies.

Cluster ensembles in three qualitatively different application scenarios:

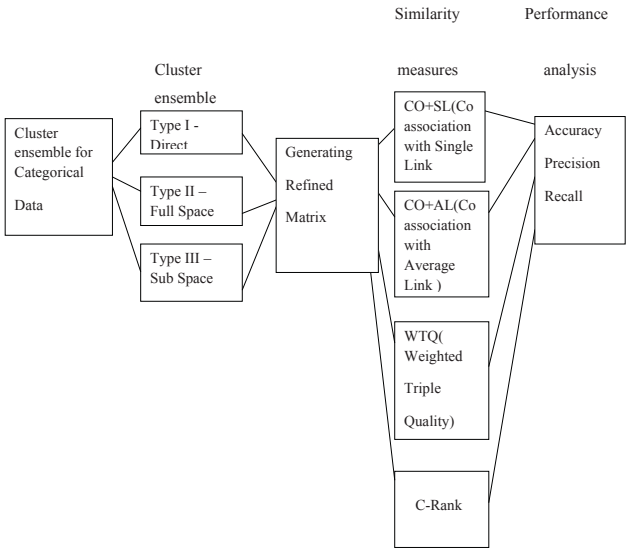
i) where the original clusters were formed based on non-identical sets of features,
ii) where the original clustering algorithms worked on non-identical sets of objects, and
iii) where a common data-set is used and the main purpose of combining multiple clustering's is to improve the quality and robustness of the solution.

The main goal of ensembles has been to improve the accuracy and robustness of a given classification or regression task, and spectacular improvements have been obtained for a wide variety of data sets. The cluster ensemble design problem is more difficult than designing classier ensembles since cluster labels are symbolic and so one must also solve a correspondence problem.

2.4 Module Description

1. Cluster Ensembles of Categorical Data
2. Creating a Cluster Ensemble
3. Generating a Refined Matrix
4. New Link-Based Similarity Algorithm
5. C-Rank link-based similarity technique

Block Diagram of the Module:



2.4.1 Cluster Ensembles of Categorical Data :

A cluster ensemble consists of different partitions. Such partitions can be obtained from multiple applications of any single algorithm with different initializations, or from the application of different algorithms to the same dataset. Cluster ensembles offer a solution to challenges inherent to clustering arising from its ill-posed nature: they can provide more robust and stable solutions by leveraging

the consensus across multiple clustering results, while averaging out emergent spurious structures that arise due to the various biases to which each participating algorithm is tuned.

2.4.2 Creating a Cluster Ensemble :

Clustering ensembles have emerged as a powerful method for improving both the robustness and the stability of unsupervised classification solutions. However, finding a consensus clustering from multiple partitions is a difficult problem that can be approached from graph-based, combinatorial or statistical perspectives. Clustering ensembles can also be used in multi objective clustering as a compromise between individual clusterings with conflicting objective functions and play an important role in distributed data mining.

Type I (Direct ensemble) :

The type of cluster ensemble transforms the problem of categorical data clustering to cluster ensembles by considering each categorical attribute value (or label) as a cluster in an ensemble. Let $X = \{ X_1 \ldots X_n\}$ be a set of N data points, $A = \{A_1 \ldots A_m\}$ be a set of categorical attributes, and $\pi = \{\pi_1, \ldots, \pi_m\}$ be a set of M partitions. Each partition π_i is generated for a specific categorical attribute $a_i \in A$.

Type II (Full-space ensemble) :

In this two ensemble types are created from base clustering results, each of which is obtained by applying a clustering algorithm to the categorical data set. In particular to a full-space ensemble, base clusterings are created from the original data, i.e., with all data attributes. To introduce an artificial instability to k-modes, the following two schemes are employed to select the number of clusters in each

base clusterings: 1) Fixed-k, k= ⌐√N¬ (where N is the no of data points), and 2) Random-k, k ∈{2,…….. ⌐√N¬ }[5].

Type III: Subspace ensemble

Another alternative to generate diversity within an ensemble is to exploit a number of different data subsets. To this extent, the cluster ensemble is established on various data subspaces, from which base clustering results are generated. Similar to the study in, for a given N *d data set of N data points and d attributes, an N * q data subspace (where q < d) is generated by $q = q_{min} + \lfloor \alpha(q_{max} - q_{min}) \rfloor$ .where $\alpha \in [0,1]$is a uniform random variable, q_{min} and q_{max} are the lower and upper bounds of the generated subspace, respectively. In particular, q_{min} and q_{max} are set to 0:75d and 0:85d.

2.4.3 Generating a Refined Matrix :

Generating a refined cluster-association matrix (RM) using a link-based similarity algorithm. Cluster ensemble methods are based on the binary cluster-association matrix. Refined cluster-association matrix is put forward as the enhanced variation of the original BM. Its aim is to approximate the value of unknown associations ("0") from known ones ("1"), whose association degrees are preserved within the RM. These hidden or unknown associations can be estimated from the similarity among clusters, discovered from a network of clusters.

$RM(x_i,c_l) = \{1, \text{ if } c_l = C_t(x_i),$

$\{ Sim(c_l, C_t(x_i)), \text{ otherwise}$

where $C_t(xi)$ is a cluster label (corresponding to a particular cluster of the clustering π_t) to which data point xi belongs. In addition, $sim(C_x,C_y) \in [0,1]$ denotes the

similarity between any two clusters C_x,C_y, which can be discovered using the following link-based algorithm.

2.4.4 New Link-Based Similarity Algorithm :

The Weighted Triple-Quality algorithm is efficient approximation of the similarity between clusters in a link network. WTQ aims to differentiate the significance of triples and hence their contributions toward the underlying similarity measure which is shown in Eqn (2.2). A cluster ensemble of a set of data points X, a weighted graph G =(V,M) can be constructed, where V is the set of vertices each representing a cluster and W is a set of weighted edges between clusters which is shown in Eqn (2.1) .

$$W_{xy} = \frac{|L_x \cap L_y|}{|L_x \cup L_y|} \quad (2.1)$$

Following that, the similarity between clusters C_x and C_y can be estimated by,

$$Sim(C_x,C_y) = \frac{WTQ_{xy}}{WTQ_{max}} * DC \quad (2.2)$$

2.3.5 Connector-based Similarity Measure

C-Rank uses both in-links and out-links at the same time. C-Rank is defined iteratively. C-Rank achieves a higher effectiveness than existing similarity measures in most cases. C-Rank converges at the 9-th iteration. When C is low, the

recursive power of C-Rank is weakened such that only the papers in local or near-local neighbourhood are used in similarity computation. When C is high, more papers in a more global neighbourhood can be used in computing the similarity recursively. When C is high, therefore, the convergence takes more time.

C-Rank link based algorithm Explanation:

C-Rank link-based algorithm is used to improve clustering quality and ranking clusters in weighted networks. C-Rank consists of three major phases: (1) identification of candidate clusters; (2) ranking the candidates by integrated cohesion; and (3) elimination of non-maximal clusters.

CHAPTER 3

RESULTS AND ANALYSIS

3.1 Experimentation

The proposed work is implemented using java. The language used for the calculation is java. The dataset generated are stored in sql server database.

3.2 System Specification

3.2.1 Hardware Requirements:

- Processor : Pentium IV
- Speed :Above 500 MHz
- RAM capacity :2 GB
- Hard disk drive :80 GB
- Key Board :Samsung 108 keys
- Mouse :Logitech Optical Mouse
- Printer : DeskJet HP
- Motherboard :Intel
- Monitor :17" Samsung

3.2.2 Software Requirements:

- Operating System :Windows XP and above

- Front end used :Java
- Back end :Sql server 2000

3.3 Software Description

U.C. I Datasets Description

The data sets used in the project are taken from the U.C.I Machine learning repository. The UCI repository is database of 177 data sets taken from various field of applications like life, computer sciences, engineering, games and social science. The attributes are of the type categorical, numerical or both. The datasets used are Breast Cancer from life sciences and Primary Tumour from the life Science area.

Breast Cancer:

Breast cancer database is collection of 400 samples. The following are the information available about this dataset:
Dataset Characteristics: Multivariate
Attribute Characteristics: Integer
Associated Tasks: Classification
Number of samples: 1200
Number of Attributes: 13
Number of Classes: 10
Area: Life Science

Primary Tumour:

Primary Tumour database is the description of the attributes about this special kind of bacteria. The following are the dataset description available:
Dataset Characteristics: Multivariate

Attribute Characteristics: Categorical
Associated Tasks: Classification
Number of samples: 800
Number of Attributes: 12
Number of Classes: 8
Area: Life

Front End : Java

Java is a high-level object-oriented programming language developed by the Sun Microsystems. Though it is associated with the World Wide Web but it is older than the origin of Web. Java is an object oriented language and a very simple language.

Java Features

Java is a set of several computer software products and specifications from Sun Microsystems (which has since merged with Oracle Corporation), that together provide a system for developing application software and deploying it in a cross-platform computing environment. Java is used in a wide variety of computing platforms from embedded devices and mobile phones on the low end, to enterprise servers and supercomputers on the high end.

Sql Server – An Overview

Microsoft SQL Server is a relational database management system developed by Microsoft. As a database, it is a software product whose primary function is to store and retrieve data as requested by other software applications, be it those on the same computer or those running on another computer across a network (including the Internet).

3.4 Problem Definition

The classification of the cluster is based on similarity measures for link based cluster approach with data pre processing module, create a cluster ensemble for categorical data module, cluster ensemble module, generating refined matrix module, Link cluster ensemble (Weighted Triple quality) module, performance module which are clearly analyzed in this chapter.

For testing the validity of the proposed model datasets were selected from the UCI Machine Learning Repository. The data sets were pre processed using classification accuracy, precision and recall. The selection of the cluster ensemble may have a significant effect on the results of the conventional methods viz., CO+SL(Co association with single link), CO+AL(Co association with average link), WTQ(Weighted Triple Quality) and the proposed methods. namely ,C-Rank. So, all the models were studied using three different cluster ensemble. Type I cluster ensemble is direct, Type II cluster ensemble is full space and Type III cluster ensemble is subspace. The parameters for the cluster ensemble were obtained after a preliminary set of experiments. The application of C-Rank in the proposed model creates a significant improvement in the accuracy level and reduction in computational time apart from getting consistent results. All the aforesaid details have been narrated in detail in this chapter.

3.5 Data Normalization

A summary of the datasets taken from the UCI Machine learning repository is shown in Table 3.1. The datasets are selected in such a way that the problems chosen are with at least six classes and no missing values.

Table 3.1 Summary of Datasets

Datasets	Number of Instances	Number of Attributes	Number of Classes	Missing Values	Area
Breast Cancer	1484	12	8	NIL	Life
Primary Tumour	699	10	10	NIL	Life

Illustration 1:

Breast Cancer Dataset:

Accuracy

Accuracy is the degree of conformity with a standard or a measure of closeness to a true value. Accuracy relates to the quality of the result obtained when compared to the standard. Accuracy is the degree of veracity while in some contexts precision may mean the degree of reproducibility which is shown in Eqn (3.1). Accuracy is dependent on how data is collected, and is usually judged by comparing several measurements from the same or different sources. The classification accuracy A_i of an individual program i depends on the number of samples correctly classified (true positives plus true negatives) and is evaluated by the formula:

$$A_i = \frac{t}{n} \times 100 \quad (3.1)$$

where

t is the number of sample correctly classified
n is the total number of sample.

The classification accuracy of standard methods(CO+SL,CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 200. If the number of cluster is 7 then type I,II,III cluster ensemble for proposed method (C-Rank) gets increased in their classification accuracy when compared to other standard methods(CO+SL,CO+AL,WTQ) are shown in the Table 3.2. The classification accuracy of standard methods(CO+SL,CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 300. If the number of cluster is 7 then type I,II,III cluster ensemble for proposed method (C-Rank) gets increased in their classification accuracy when compared to other standard methods(CO+SL,CO+AL,WTQ) are shown in the Table 3.3. The classification accuracy of standard methods(CO+SL,CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 400. If the number of cluster is 7 then type I,II,III cluster ensemble for proposed method (C-Rank) gets increased in their classification accuracy when compared to other standard methods(CO+SL,CO+AL,WTQ) are shown in the Table 3.4.

Table 3.2: Comparison of Classification Accuracy of standard and proposed methods based on number of samples = 200.

Number of Cluster	Ensemble Type	Classification Accuracy (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C -Rank
3	Type I	68.75	69.84	85.06	92.09
	Type II	73.93	75.84	86.85	92.64
	Type III	79.66	80.20	87.78	92.76
4	Type I	68.96	69.91	85.12	93.66
	Type II	73.94	74.85	85.60	94.45
	Type III	79.65	84.66	90.77	94.56

5	Type I	68.80	69.88	84.67	93.51
	Type II	74.21	76.65	86.39	95.30
	Type III	77.23	83.78	88.37	95.43
6	Type I	68.67	69.72	84.56	94.34
	Type II	74.77	79.72	87.89	95.48
	Type III	82.56	88.45	91.42	95.79

Table 3.3: Comparison of Classification Accuracy of standard and proposed methods based on number of samples = 300.

Number of Cluster	Ensemble Type	Classification Accuracy (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C -Rank
3	Type I	70.93	71.05	87.59	93.35
	Type II	75.35	77.29	88.01	93.54
	Type III	81.89	81.90	89.43	94.49
4	Type I	70.95	71.28	87.62	94.29
	Type II	75.68	77.33	87.75	96.67
	Type III	81.87	85.76	91.76	96.72
5	Type I	70.97	71.22	86.72	94.82
	Type II	76.56	78.89	88.18	96.75
	Type III	79.45	85.34	90.76	96.83
6	Type I	70.92	71.60	86.84	95.78
	Type II	76.87	81.33	89.33	97.73
	Type III	84.29	89.37	92.83	97.89

7	Type I	70.21	72.55	87.72	95.61
	Type II	83.88	88.78	92.77	97.77
	Type III	84.67	89.56	93.56	99.20

Table 3.4: Comparison of Classification Accuracy of standard and proposed methods based on number of samples = 400.

Number of Cluster	Ensemble Type	Classification Accuracy (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C -Rank
3	Type I	72.65	73.93	88.98	94.98
	Type II	77.31	79.89	89.99	95.44
	Type III	83.62	84.39	89.97	95.85
4	Type I	72.66	73.96	89.96	95.44
	Type II	77.35	79.92	89.99	97.77
	Type III	82.89	86.89	92.80	97.96
5	Type I	72.71	74.09	88.89	95.88
	Type II	78.88	80.56	90.89	97.92
	Type III	80.96	86.78	92.67	97.98
6	Type I	72.48	73.82	89.91	96.50
	Type II	78.92	83.66	91.88	97.94
	Type III	85.99	90.67	93.44	98.92
7	Type I	72.77	75.20	89.90	96.84
	Type II	84.90	90.22	93.97	98.90
	Type III	85.89	90.99	94.78	99.64

Precision:

Precision is the degree of refinement in the performance of an operation (procedures and instrumentation) or in the statement of a result.

$$\text{Precision}(i,j) = n_{ij} / n_j \qquad (3.2)$$

where,

n_{ij} = number of member of class i in cluster j.

n_j = number of members of cluster j.

The Precision of standard methods (CO+SL,CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 200. If the number of cluster is 7 then type I,II,III cluster ensemble for proposed method (C-Rank) gets increased in their Precision value when compared to other standard methods(CO+SL,CO+AL,WTQ)are shown in the Table 3.5. The Precision of standard methods (CO+SL,CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 300 which is shown in Eqn (3.2). If the number of cluster is 7 then type I,II,III cluster ensemble for proposed method (C-Rank) gets increased in their Precision value when compared to other standard methods(CO+SL,CO+AL,WTQ)are shown in the Table 3.6. The Precision of standard methods (CO+SL,CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 400. If the number of cluster is 7 then type I,II,III cluster ensemble for proposed method (C-Rank) gets increased in their Precision value when compared to other standard methods(CO+SL,CO+AL,WTQ)are shown in the Table 3.7.

Table 3.5: Comparison of Precision of standard and proposed methods based on number of samples = 200.

Number of Cluster	Ensemble Type	Precision (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C -Rank
3	Type I	68.77	69.88	85.08	92.67
	Type II	73.93	75.87	87.87	92.11
	Type III	79.97	79.35	86.79	92.78
4	Type I	68.96	69.43	85.18	93.67
	Type II	73.94	74.89	85.60	94.46
	Type III	79.67	84.68	90.77	94.58
5	Type I	68.83	69.88	84.68	93.53
	Type II	76.58	78.89	88.22	96.89
	Type III	77.25	83.78	88.39	96.90
6	Type I	68.68	69.97	84.58	94.37
	Type II	74.69	79.72	87.89	95.78
	Type III	82.59	88.59	91.45	95.80
7	Type I	68.58	70.72	85.35	94.46
	Type II	81.49	86.44	91.77	96.48
	Type III	82.79	88.42	92.89	97.87

Table 3.6: Comparison of Precision of standard and proposed methods based on number of samples = 300.

Number of Cluster	Ensemble Type	Precision (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C- Rank
3	Type I	70.96	71.09	87.67	93.37
	Type II	75.36	77.29	89.05	93.58
	Type III	81.89	81.73	88.45	94.52
4	Type I	70.93	71.35	87.65	94.30
	Type II	75.42	77.33	87.73	96.35
	Type III	81.87	85.77	91.77	96.58
5	Type I	70.93	71.25	86.74	94.83
	Type II	74.24	76.67	86.77	95.34
	Type III	79.47	85.39	90.74	96.73
6	Type I	70.91	71.63	86.82	95.79
	Type II	76.88	81.39	89.37	96.35
	Type III	84.32	89.38	92.84	97.90
7	Type I	70.25	72.59	87.76	95.63
	Type II	83.89	88.79	92.78	97.78
	Type III	84.69	89.58	93.59	99.24

Table 3.7: Comparison of Precision of standard and proposed methods based on number of samples = 400.

Number of Cluster	Ensemble Type	Precision(%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C- Rank
3	Type I	72.67	73.95	88.99	94.98
	Type II	77.33	79.89	89.99	95.46
	Type III	83.66	84.41	91.85	95.85
4	Type I	72.68	74.96	89.97	95.47
	Type II	77.39	79.91	89.98	97.73
	Type III	82.89	86.89	92.80	97.88
5	Type I	72.73	74.12	88.89	95.88
	Type II	78.88	80.58	90.89	97.92
	Type III	80.96	86.78	92.87	97.99
6	Type I	73.48	75.84	89.91	96.50
	Type II	78.94	83.63	91.88	97.94
	Type III	85.99	90.67	93.93	98.93
7	Type I	73.78	76.20	89.92	96.89
	Type II	84.93	90.25	93.97	98.91
	Type III	85.91	90.99	94.78	99.65

Recall Rate:

The recall rate is calculated as,

$$Recall(i,j) = n_{ij} / n_i \qquad (3.3)$$

where,

n_{ij} = number of members of class i in cluster j.

n_i = number of members of class i.

The Recall of standard methods (CO+SL,CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 300. If the no of cluster is 7 then type I,II,III cluster ensemble for proposed method (C-Rank) gets increased in their Recall value when compared to other standard methods(CO+SL,CO+AL,WTQ)are shown the Table 3.8. The Recall of standard methods (CO+SL,CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 300 which is shown in Eqn (3.3). If the no of cluster is 7 then type I,II,III cluster ensemble for proposed method (C-Rank) gets increased in their Recall value when compared to other standard methods(CO+SL,CO+AL,WTQ)are shown the Table 3.9. The Recall of standard methods (CO+SL, CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 400. If the number of cluster is 7 then type I,II,III cluster ensemble for proposed method (C-Rank) gets increased in their Recall value when compared to other standard methods(CO+SL,CO+AL,WTQ)are shown in the Table 3.10.

Table 3.8: Comparison of Recall of standard and proposed methods based on number of samples = 200.

Number of Cluster	Ensemble Type	Recall (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C -Rank
3	Type I	68.78	69.89	85.08	92.70
	Type II	73.93	75.87	87.87	92.19
	Type III	79.99	79.38	86.80	92.79
4	Type I	68.96	69.47	85.22	93.69
	Type II	73.94	74.90	85.60	94.48
	Type III	79.71	84.72	90.77	94.60
5	Type I	68.87	69.89	84.71	93.58
	Type II	76.62	78.90	88.25	95.90
	Type III	77.28	83.79	88.44	95.92
6	Type I	68.68	69.97	84.59	94.41
	Type II	74.69	79.75	87.89	96.79
	Type III	82.59	88.67	91.48	96.83
7	Type I	68.58	70.77	85.38	94.48
	Type II	81.56	86.44	91.79	96.49
	Type III	82.82	88.46	92.91	97.88

Table 3.9: Comparison of Recall of standard and proposed methods based on number of samples = 300.

Number of Cluster	Ensemble Type	Recall (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C -Rank
3	Type I	71.96	73.09	87.69	93.39
	Type II	75.39	77.34	89.09	93.60
	Type III	81.89	81.75	88.47	94.55
4	Type I	72.93	73.35	87.65	94.38
	Type II	75.42	77.37	87.76	96.39
	Type III	81.87	85.79	91.78	96.58
5	Type I	72.93	74.28	86.74	94.83
	Type II	74.28	76.71	86.79	95.34
	Type III	79.51	85.42	90.76	96.78
6	Type I	73.94	75.66	86.85	95.79
	Type II	76.89	81.45	89.40	96.37
	Type III	84.35	89.41	92.87	97.92
7	Type I	74.27	76.64	87.79	95.66
	Type II	83.91	88.81	92.87	97.80
	Type III	84.72	89.60	93.65	99.27

Table 3.10: Comparison of Recall of standard and proposed methods based on number of samples = 400.

Number of Cluster	Ensemble Type	Recall (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C- Rank
3	Type I	73.67	73.95	88.99	94.98
	Type II	77.37	79.89	89.99	95.46
	Type III	83.70	84.41	91.88	95.85
4	Type I	73.71	74.96	89.98	95.47
	Type II	77.42	79.95	89.99	97.73
	Type III	82.91	86.91	92.85	97.88
5	Type I	74.75	74.15	88.92	95.88
	Type II	78.89	82.66	90.93	97.97
	Type III	81.96	86.78	92.95	97.99
6	Type I	74.56	75.84	89.97	96.55
	Type II	79.94	83.65	91.93	97.96
	Type III	86.99	91.67	93.95	98.97
7	Type I	74.78	76.27	89.94	96.89
	Type II	84.93	90.56	93.98	98.95
	Type III	87.91	91.99	94.80	99.68

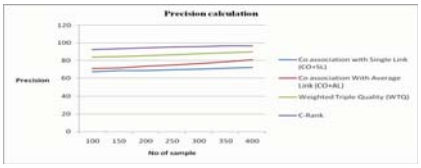


Fig.3.11 Graph for Performance of precision based on number of samples

The above graph in the Fig.3.11 shows that if number of sample are more then precision value for proposed methods(C-Rank) has increased up to 97.76% . The precision value for standard methods(CO+SL,CO+AL,WTQ) are slightly less when compared to proposed methods.

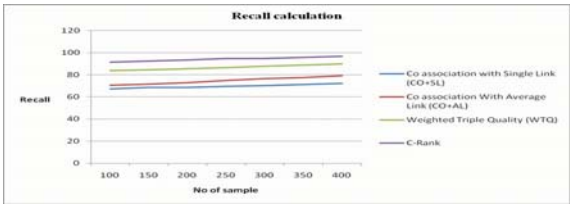


Fig.3.12 Graph for Performance of Recall rate based on number of samples

The above graph in the Fig.3.12 shows that if number of sample are more then recall value for proposed methods(C-Rank) has increased up to 96.76% . The recall value for standard methods(CO+SL,CO+AL,WTQ) are slightly less when compared to proposed methods.

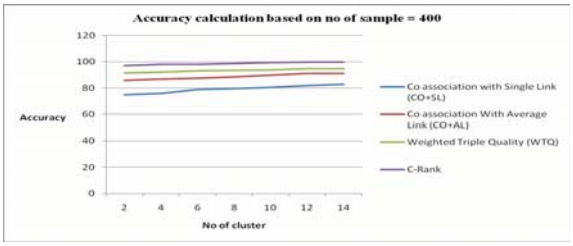


Fig.3.13 Graph for Performance of Accuracy based on number of sample = 400.

The above graph in the Fig.3.13 shows that if number of sample are more then accuracy value for proposed methods(C-Rank) has increased up to 99.99% . The accuracy value for standard methods(CO+SL,CO+AL,WTQ) are slightly less when compared to proposed methods.

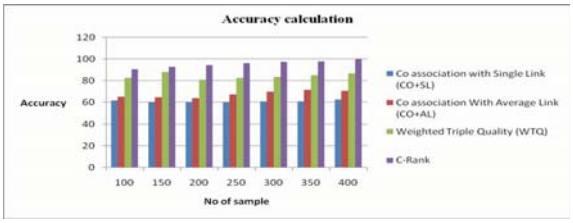


Fig.3.14 Graph for Performance of Accuracy based on number of samples

The above graph in the Fig.3.14 shows that the if number of sample is 100 then it shows the accuracy value for both proposed and standard methods in bar chart format. If number of sample is 400 then accuracy for proposed methods is 99.99 % but for standard method like WTQ has reached 90% , other standard methods are less when compared to proposed methods.

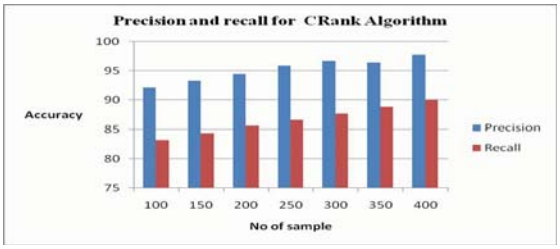


Fig.3.15 Graph for comparison of precision and recall for C-Rank algorithm

The above graph in the Fig.3.15 shows that comparison of precision and recall for C-Rank algorithm is that if number of sample are more then precision value is also gets increased when compared to recall value. If number of sample is 400 then precision value has reached 97% when compared to recall value by using C-Rank algorithm.

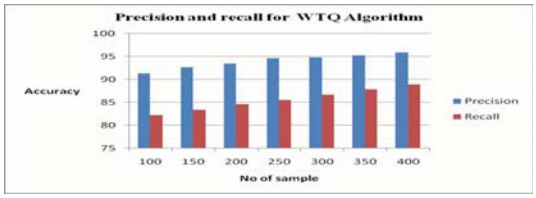


Fig.3.16 Graph for comparison of precision and recall for WTQ algorithm

If the number of sample are more then precision value is also gets increased when compared to recall value. If number of sample is 400 then precision value has reached 96% when compared to recall value by using WTQ algorithm which is in the Fig.3.16.

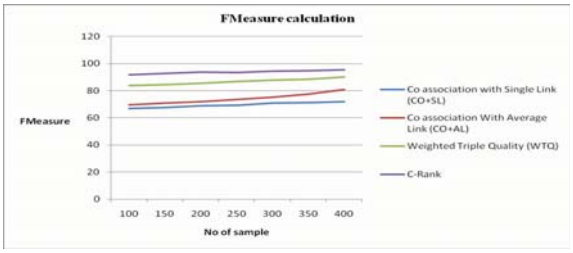


Fig.3.17 Graph for Performance of FMeasure based on number of samples

If the number of sample are more then FMeasure value for proposed methods(C-Rank) has increased up to 94.95% . The FMeasure value for standard

methods(CO+SL,CO+AL,WTQ) are slightly less when compared to proposed methods which is shown in the Fig.3.17.

Illustration 2:

Primary Tumour Datasets

Accuracy:

The classification accuracy of standard methods (CO+SL,CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 200. If the number of cluster is 7 then type I,II,III cluster ensemble for proposed method (C-Rank) gets increased in their classification accuracy when compared to other standard methods(CO+SL,CO+AL,WTQ)are shown the Table 3.18. The classification accuracy of standard methods (CO+SL,CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 300 which is shown in Eqn (3.4). If the no of cluster is 7 then type I,II,III cluster ensemble for proposed method (C-Rank) gets increased in their classification accuracy when compared to other standard methods(CO+SL,CO+AL,WTQ) are shown in the Table 3.19. The classification accuracy of standard methods (CO+SL,CO+AL and WTQ) and proposed method (C-Rank) based on number of samples is 400. If the number of cluster is 7 then type I,II,III cluster ensemble for proposed method (C-Rank) gets increased in their classification accuracy when compared to other standard methods(CO+SL,CO+AL,WTQ)are shown the Table 3.20.

Table 3.18: Comparison of Classification Accuracy of standard and proposed methods based on number of samples = 200.

Number of Cluster	Ensemble Type	Classification Accuracy (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C- Rank
3	Type I	31.48	39.47	41.67	44.57
	Type II	33.28	40.63	40.78	50.35
	Type III	35.34	42.48	44.35	48.55
4	Type I	31.40	38.77	41.56	49.53
	Type II	32.68	40.34	43.77	50.42
	Type III	35.37	42.56	44.82	51.03
5	Type I	31.60	39.50	41.55	49.60
	Type II	33.30	40.67	43.24	50.37
	Type III	35.38	42.50	44.62	48.73
6	Type I	31.50	39.54	41.57	48.63
	Type II	33.35	40.69	43.34	49.41
	Type III	35.41	42.58	44.56	50.43
7	Type I	31.58	39.57	41.54	48.69
	Type II	33.73	40.69	43.56	49.71
	Type III	35.67	42.55	44.58	50.85

Table 3.19: Comparison of Classification Accuracy of standard and proposed methods based on number of samples = 300.

Number of Cluster	Ensemble Type	Classification Accuracy (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C -Rank
3	Type I	32.45	40.68	44.56	53.65
	Type II	37.48	44.20	45.68	53.88
	Type III	38.59	45.40	47.59	54.06
4	Type I	32.52	40.17	44.53	53.69
	Type II	35.78	43.45	47.24	53.90
	Type III	38.54	46.67	47.83	54.22
5	Type I	32.88	40.55	44.80	54.40
	Type II	37.55	44.34	47.50	54.69
	Type III	38.60	46.70	47.77	55.10
6	Type I	32.98	40.78	44.93	54.88
	Type II	37.74	44.89	47.67	55.37
	Type III	39.01	46.91	47.89	55.94
7	Type I	33.10	41.45	45.39	56.44
	Type II	38.22	45.67	48.45	56.77
	Type III	39.56	46.98	49.56	57.20

Table 3.20: Comparison of Classification Accuracy of standard and proposed methods based on number of samples = 400.

Number of Cluster	Ensemble Type	Classification Accuracy (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C- Rank
3	Type I	38.10	42.74	46.12	56.22
	Type II	39.64	46.40	47.53	56.56
	Type III	41.73	47.22	48.32	57.02
4	Type I	38.36	42.89	46.45	56.30
	Type II	39.78	46.67	47.60	56.67
	Type III	41.84	47.46	48.56	57.22
5	Type I	38.56	43.03	46.67	56.59
	Type II	40.29	46.78	47.73	56.77
	Type III	42.02	47.88	48.60	57.30
6	Type I	38.88	43.49	46.72	56.69
	Type II	40.56	46.89	47.84	56.80
	Type III	42.34	47.90	48.68	57.60

Precision:

Table 3.21: Comparison of Precision of standard and proposed methods based on number of samples = 200.

Number of Cluster	Ensemble Type	Precision (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C- Rank
3	Type I	31.49	39.50	41.69	44.60
	Type II	33.34	40.67	40.81	50.44
	Type III	35.39	42.52	44.39	48.59
4	Type I	31.45	38.79	41.70	49.58
	Type II	32.71	40.38	43.81	50.45
	Type III	35.42	42.66	44.85	51.09
5	Type I	31.63	39.58	41.61	49.64
	Type II	33.36	40.69	43.29	50.39
	Type III	35.41	42.54	44.67	48.76
6	Type I	31.53	39.59	41.61	48.69
	Type II	33.37	40.73	43.38	49.46
	Type III	35.46	42.64	44.68	50.49
7	Type I	31.63	39.66	41.59	48.73
	Type II	33.79	40.72	43.66	49.76
	Type III	35.72	42.59	44.72	50.88

Table 3.22: Comparison of Precision of standard and proposed methods based on number of samples = 300.

Number of Cluster	Ensemble Type	Precision (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C -Rank
3	Type I	32.48	40.70	44.59	53.69
	Type II	37.51	44.24	45.70	53.90
	Type III	38.62	45.47	47.62	54.12
4	Type I	32.55	40.24	44.56	53.71
	Type II	35.80	43.49	47.26	53.93
	Type III	38.58	46.77	47.89	54.27
5	Type I	32.89	40.59	44.82	54.45
	Type II	37.58	44.38	47.54	54.72
	Type III	38.64	46.74	47.79	55.19
6	Type I	33.02	40.80	44.95	54.89
	Type II	37.79	44.91	47.77	55.42
	Type III	39.09	46.93	47.90	56.04
7	Type I	33.14	41.48	45.44	56.51
	Type II	38.26	45.69	48.48	56.79
	Type III	39.60	47.02	49.61	57.25

Table 3.23: Comparison of Precision of standard and proposed methods based on number of samples = 400.

Number of Cluster	Ensemble Type	Precision (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C- Rank
3	Type I	38.14	42.79	46.14	56.26
	Type II	39.68	46.45	47.59	56.60
	Type III	41.76	47.29	48.42	57.07
4	Type I	38.39	42.90	46.49	56.34
	Type II	39.80	46.69	47.65	56.71
	Type III	41.87	47.51	48.59	57.26
5	Type I	38.59	43.11	46.69	56.62
	Type II	40.33	46.80	47.76	56.79
	Type III	42.09	47.91	48.64	57.38
6	Type I	38.89	43.54	46.77	56.76
	Type II	40.63	46.91	47.88	56.83
	Type III	42.38	47.94	48.72	57.65
7	Type I	39.18	43.69	46.87	56.91
	Type II	40.81	46.93	47.95	57.74
	Type III	42.73	47.95	48.79	57.82

Recall Rate :

Table 3.24: Comparison of Recall of standard and proposed methods based on number of samples = 200.

Number of Cluster	Ensemble Type	Recall (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C -Rank
3	Type I	31.52	39.56	41.71	44.64
	Type II	33.36	40.69	40.85	50.48
	Type III	35.44	42.57	44.47	48.63
4	Type I	31.49	38.80	41.75	49.63
	Type II	32.74	40.42	43.86	50.49
	Type III	35.45	42.69	44.89	51.23
5	Type I	31.67	39.64	41.68	49.72
	Type II	33.39	40.73	43.33	50.45
	Type III	35.45	42.58	44.69	48.80
6	Type I	31.58	39.59	41.61	48.69
	Type II	33.40	40.73	43.38	49.46
	Type III	35.52	42.64	44.68	50.49
7	Type I	31.68	39.71	41.62	48.78
	Type II	33.82	40.78	43.69	49.83
	Type III	35.77	42.63	44.76	50.91

Table 3.25: Comparison of Recall of standard and proposed methods based on number of samples = 300.

Number of Cluster	Ensemble Type	Recall (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C- Rank
3	Type I	32.53	40.75	44.63	53.73
	Type II	37.58	44.29	45.74	53.93
	Type III	38.67	45.56	47.67	54.18
4	Type I	32.59	40.29	44.64	53.76
	Type II	35.84	43.55	47.29	53.97
	Type III	38.63	46.82	47.92	54.33
5	Type I	32.91	40.62	44.86	54.48
	Type II	37.64	44.43	47.59	54.76
	Type III	38.72	46.79	47.83	55.23
6	Type I	33.07	40.83	44.98	54.91
	Type II	37.82	44.94	47.81	55.46
	Type III	39.13	46.97	47.94	56.12
7	Type I	33.20	41.53	45.49	56.58
	Type II	38.34	45.72	48.56	56.83
	Type III	39.65	47.07	49.68	57.29

Table 3.26: Comparison of Recall of standard and proposed methods based on number of samples = 400.

Number of Cluster	Ensemble Type	Recall (%)			
		Co association with Single Link (CO+SL)	Co association with Average Link (CO+AL)	Weighted Triple Quality (WTQ)	C- Rank
3	Type I	38.19	42.88	46.20	56.29
	Type II	39.76	46.49	47.63	56.68
	Type III	41.83	47.34	48.52	57.16
4	Type I	38.47	42.98	46.53	56.40
	Type II	39.86	46.74	47.69	56.77
	Type III	41.94	47.58	48.75	57.28
5	Type I	38.67	43.16	46.72	56.65
	Type II	40.39	46.85	47.79	56.82
	Type III	42.23	47.97	48.68	57.45
6	Type I	38.93	43.67	46.83	56.79
	Type II	40.68	46.96	47.93	56.87
	Type III	42.45	47.99	48.78	57.84
7	Type I	39.25	43.77	46.93	56.97
	Type II	40.86	46.98	47.99	57.79
	Type III	42.79	47.99	48.86	57.88

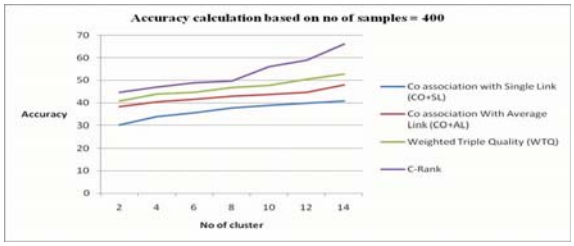


Fig.3.27 Graph for Performance of Accuracy based on number of sample = 400.

If the number of clusters are more then accuracy value for proposed methods(C-Rank) has increased up to 65.48% . The accuracy value for standard methods(CO+SL,CO+AL,WTQ) are slightly less when compared to proposed methods which is shown in the Fig.3.27.

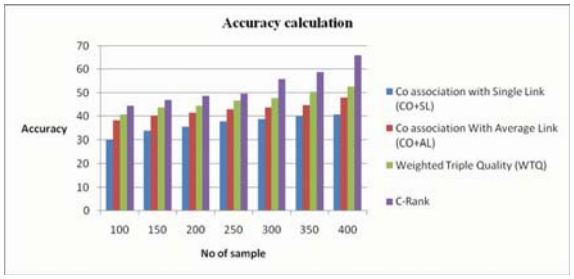


Fig.3.28 Graph for Performance of Accuracy based on number of samples

If the number of sample is 100 then it shows the accuracy value for both proposed and standard methods in bar chart format. If number of sample is 400 then accuracy for proposed methods is 68.66 % but for standard method like WTQ has reached 53.80 % , other standard methods are less when compared to proposed methods which is shown in the Fig.3.28.

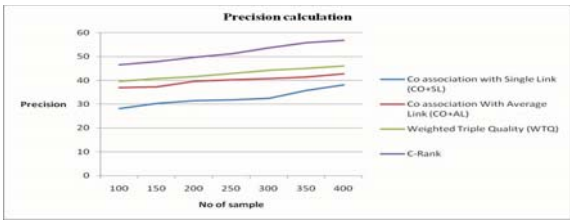


Fig.3.29 Graph for Performance of Precision based on number of samples

If the number of sample are more then precision value for proposed methods(C-Rank) has increased up to 58.79 % . The precision rate for standard methods(CO+SL,CO+AL,WTQ) are slightly less when compared to proposed methods which is shown in the Fig.3.29.

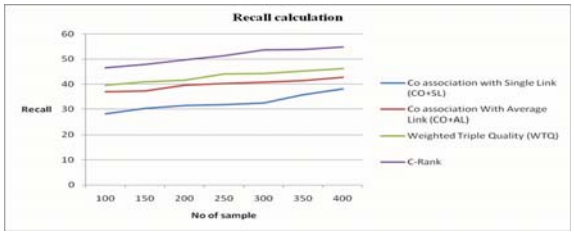


Fig.3.30 Graph for Performance of Recall based on number of samples

If the number of sample are more then recall value for proposed methods(C-Rank) has increased up to 54.60 % . The recall rate for standard methods(CO+SL,CO+AL,WTQ) are slightly less when compared to proposed methods which is shown in the Fig.3.30.

.

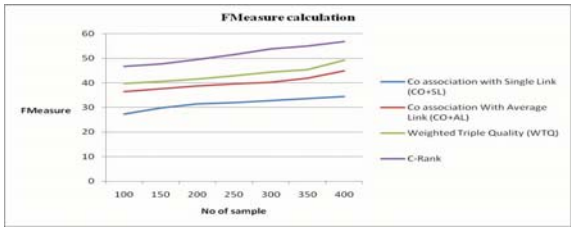


Fig.3.31 Graph for Performance of FMeasure based on number of sample

If the number of sample are more then FMeasure value for proposed methods(C-Rank) has increased up to 56.62 % . The FMeasure value for standard methods(CO+SL,CO+AL,WTQ) are slightly less when compared to proposed methods which is shown in the Fig.3.31.

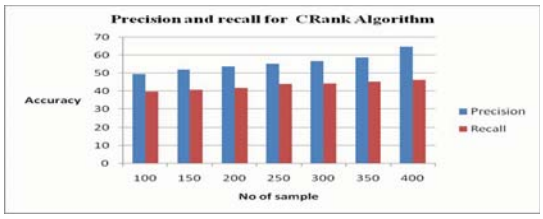


Fig.3.32 Graph for comparison of precision and recall for C-Rank algorithm

The comparison of precision and recall for C-Rank algorithm is that if number of sample are more then precision value is also gets increased when compared to recall value. If number of sample is 400 then precision value has reached 65 % when compared to recall value by using C-Rank algorithm which is shown in the Fig.3.32.

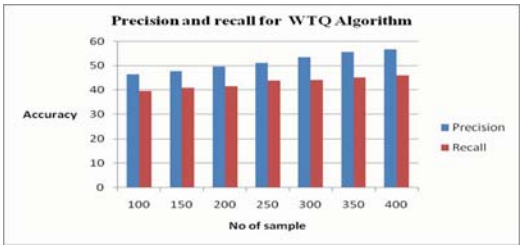


Fig.3.33 Graph for comparison of precision and recall for WTQ algorithm

The comparison of precision and recall for WTQ algorithm is that if number of sample are more then precision value is also gets increased when compared to recall value. If number of sample is 400 then precision value has reached 56% when compared to recall value by using WTQ algorithm which is shown in the Fig.3.33.

3.7 Conclusion and Future Enhancement

3.7.1 Conclusion

It presents a novel, highly effective link-based cluster ensemble approach to categorical data clustering. It transforms the original categorical data matrix to an information-preserving numerical variation (RM), to which an effective graph partitioning technique can be directly applied. The problem of constructing the RM is efficiently resolved by the similarity among categorical labels (or clusters), using the Weighted Triple-Quality similarity algorithm. The empirical study, with different ensemble types, validity measures, and data sets, suggests that the

proposed link-based method usually achieves superior clustering results compared to those of the traditional categorical data algorithms and benchmark cluster ensemble techniques. Also, the new method will be applied to specific domains, including tourism and medical data sets.

3.7.2 Future Work:

To improve clustering quality a new link-based approach the conventional matrix by discovering unknown entries through similarity between clusters in an ensemble and an efficient link-based algorithm is proposed for the underlying similarity assessment. To extend the work by analyzing the behaviour of other link-based similarity measures with this problem the quality of the clustering result. C-Rank link-based algorithm is used to improve clustering quality and ranking clusters in weighted networks. C-Rank consists of three major phases: (1) identification of candidate clusters; (2) ranking the candidates by integrated cohesion; and (3) elimination of non-maximal clusters. Finally apply this clustering result in graph partitioning technique is applied to a weighted bipartite graph that is formulated from the refined matrix.

APPENDIX

SOURCE CODE

Choose database:

```
import java.awt.BorderLayout;
import java.awt.Color;
import java.awt.Graphics;
import java.awt.event.ActionEvent;
import java.awt.event.ActionListener;
class ChooseDB extends JFrame{
   ChooseDB()throws Exception{
 setDefaultCloseOperation(JFrame.EXIT_ON_CLOSE);
 add(new ImagePanel(),BorderLayout.CENTER);
setSize(1000, 600);
setResizable(false);
 setVisible(true);
   }
   public static void main(String args[])throws Exception{
      new ChooseDB();
   } }
class ImagePanel extends JPanel{
   public BufferedImage img;
```

```
    private JComboBox list;
ImagePanel()throws Exception{
        setLayout(null);
```

Centroid:

```
class Centroid {
double mx1,mx2,mx3,mx4,mx5,mx6,mx7,mx8;
private Cluster mCluster;
public Centroid(double x1,double x2,double x3,double x4,double x5,double
x6,double x7,double x8) {
this.mx1 = x1;
this.mx2 = x2;
this.mx3 = x3;
this.mx4 = x4;
this.mx5 = x5;
this.mx6 = x6;
this.mx7 = x7;
this.mx8 = x8;   }
//caluclating the new Centroid
        for (i = 0; i < numDP; i++) {
                        tempx1=tempx1+mCluster.getDataPoint(i).getx1();
                        tempx2=tempx2+mCluster.getDataPoint(i).getx2();
                        tempx3=tempx3+mCluster.getDataPoint(i).getx3();
                        tempx4=tempx4+mCluster.getDataPoint(i).getx4();
```



```
                        tempx5=tempx5+mCluster.getDataPoint(i).getx5();
                        tempx6=tempx6+mCluster.getDataPoint(i).getx6();
                        tempx7=tempx7+mCluster.getDataPoint(i).getx7();
                        tempx8=tempx8+mCluster.getDataPoint(i).getx8();   }
```

Datapoints:

```
public class DataPoint {
private double mx1,mx2,mx3,mx4,mx5,mx6,mx7,mx8;
private String mObjName;
private Cluster mCluster;
private double mEuDt;
public DataPoint(double x1,double x2, double x3,double x4, double x5, double x6,
double x7, double x8,  String name) {
 public void calcEuclideanDistance() {
mEuDt = Math.sqrt(Math.pow((mx1 -
mCluster.getCentroid().getx1()),2)+Math.pow((mx2 -
mCluster.getCentroid().getx2()),2)+Math.pow((mx3 -
mCluster.getCentroid().getx3()),2)+Math.pow((mx4 -
mCluster.getCentroid().getx4()),2)+Math.pow((mx5 -
mCluster.getCentroid().getx5()),2)+Math.pow((mx6 -
mCluster.getCentroid().getx6()),2)+Math.pow((mx7 -
mCluster.getCentroid().getx7()),2)+Math.pow((mx8 -
mCluster.getCentroid().getx8()),2));  }
public double testEuclideanDistance(Centroid c) {
        return Math.sqrt(Math.pow((mx1 - c.getx1()),2)+Math.pow((mx2 -
c.getx2()),2)+Math.pow((mx3 - c.getx3()),2)+Math.pow((mx4 -
c.getx4()),2)+Math.pow((mx5 - c.getx5()),2)+Math.pow((mx6 -
c.getx6()),2)+Math.pow((mx7 - c.getx7()),2)+Math.pow((mx8 - c.getx8()),2) ); }
```


Cluster demo:

```
 import java.util.Vector;
 public class ClusterDemo {
 private Cluster[] clusters;
                x1 = (((getMaxx1Value() - getMinx1Value()) / (clusters.length + 1))
* n) + getMinx1Value();
                x2 = (((getMaxx2Value() - getMinx2Value()) / (clusters.length + 1))
* n) + getMinx2Value();
                x3 = (((getMaxx3Value() - getMinx3Value()) / (clusters.length + 1))
* n) + getMinx3Value();
                x4 = (((getMaxx4Value() - getMinx4Value()) / (clusters.length + 1))
* n) + getMinx4Value();
                x5 = (((getMaxx5Value() - getMinx5Value()) / (clusters.length + 1))
* n) + getMinx5Value();
                x6 = (((getMaxx6Value() - getMinx6Value()) / (clusters.length + 1))
* n) + getMinx6Value();
                x7 = (((getMaxx7Value() - getMinx7Value()) / (clusters.length + 1))
* n) + getMinx7Value();
                x8 = (((getMaxx8Value() - getMinx8Value()) / (clusters.length + 1))
* n) + getMinx8Value();
```


SCREEN SHOTS

Breast Cancer Datasets Screen shots:

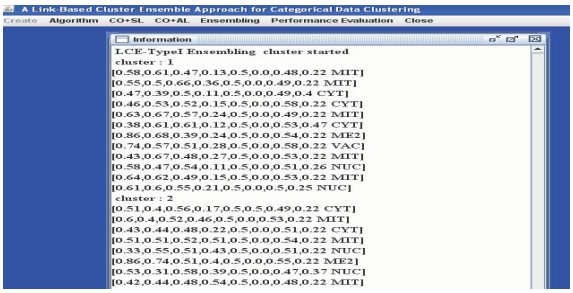


Fig A1.Type I direct cluster ensemble results for LCE algorithm

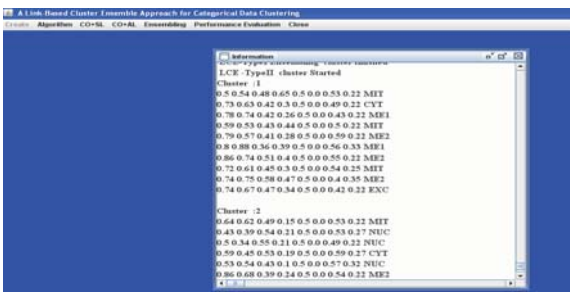


Fig A2.Type II full space cluster ensemble results for LCE algorithm

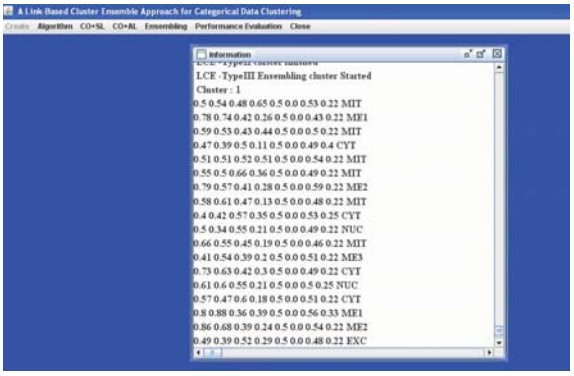


Fig A3.Type III sub space cluster ensemble results for LCE algorithm

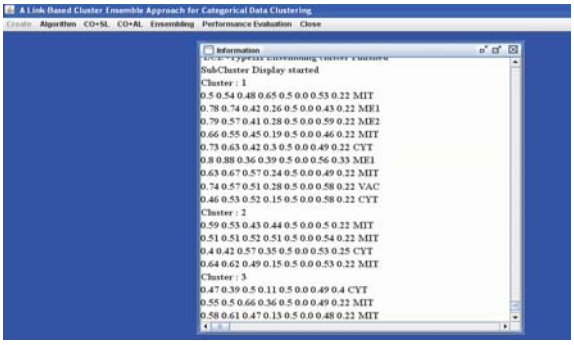


Fig A4.Refined matrix results for LCE algorithm

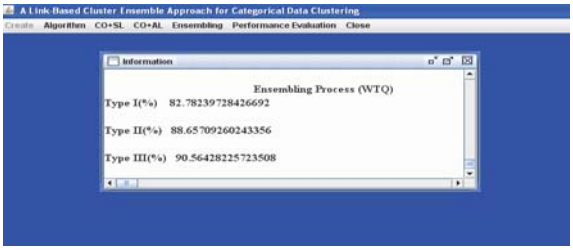


Fig A5.Performance evaluation of LCE algorithm

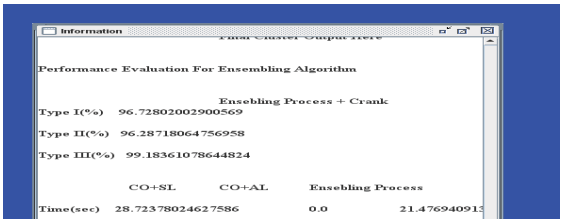


Fig A6.Performance evaluation of C-Rank algorithm

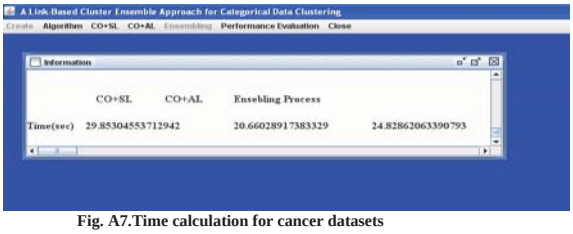


Fig. A7.Time calculation for cancer datasets

Tumour Datasets Screen Shots:

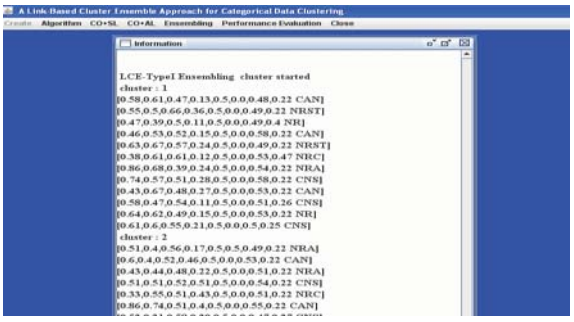


Fig A8.Type I direct cluster ensemble results for LCE algorithm

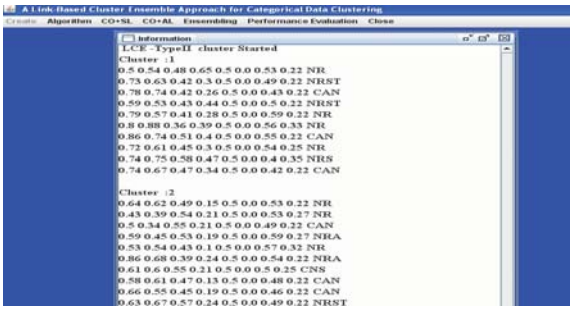


Fig A9.Type II full space cluster ensemble results for LCE algorithm

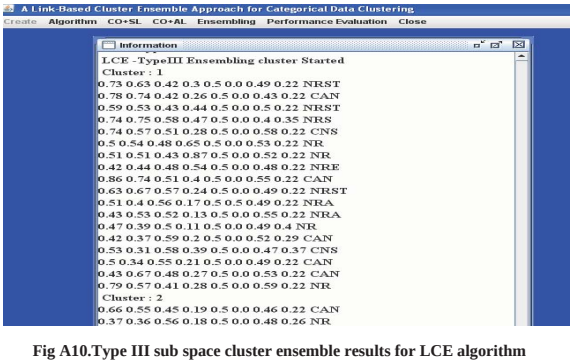


Fig A10.Type III sub space cluster ensemble results for LCE algorithm

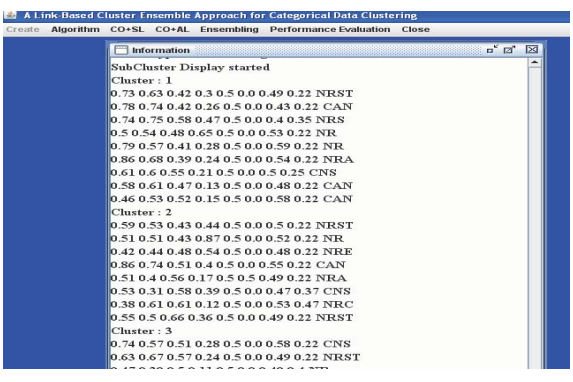


Fig A11.Refined matrix results for LCE algorithm

Fig A12.Performance evaluation of LCE algorithm

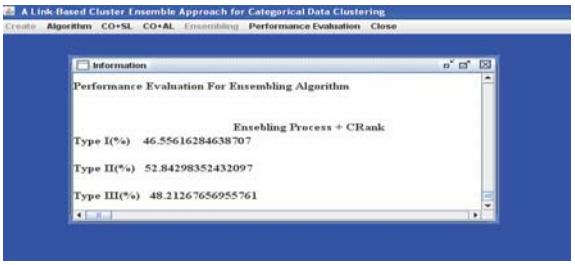


Fig A13.Performance evaluation of C-Rank algorithm

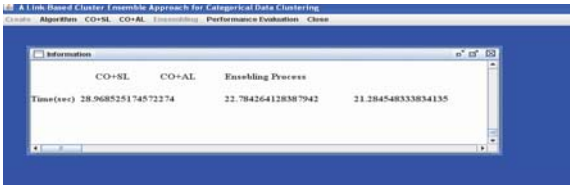


Fig A14.Time calculation for tumour dataset

REFERENCES

1. Gionis, H. Mannila, and P. Tsaparas, "Clustering Aggregation," Proc. Int'l Conf. Data Eng. (ICDE), pp. 341-352, 2005.
2. A.P. Topchy, A.K. Jain, and W.F. Punch, "Clustering Ensembles: Models of Consensus and Weak Partitions," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 12, pp. 1866-1881, Dec. 2005.
3. Boulis and M. Ostendorf, "Combining Multiple Clustering Systems," Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 63-74, 2004.
4. Domeniconi and M. Al-Razgan, "Weighted Cluster Ensembles: Methods and Analysis," ACM Trans. Knowledge Discovery from Data, vol. 2, no. 4, pp. 1-40, 2009.
5. Gibson, J. Kleinberg, and P. Raghavan, "Clustering Categorical Data: An Approach Based on Dynamical Systems," VLDB J., vol. 8, nos. 3-4, pp. 222-236, 2000.
6. G. Das, H. Mannila, and P. Ronkainen, "Similarity of Attributes by External Probes," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 16-22, 1998.
7. M. Law, A. Topchy, and A.K. Jain, "Multiobjective Data Clustering," Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 424-430, 2004.
8. M. Al-Razgan, C. Domeniconi, and D. Barbara, "Random Subspace Ensembles for Clustering Categorical Data," Supervised and Unsupervised Ensemble Methods and Their Applications, pp. 31-48, Springer, 2008.
9. A. Asuncion and D.J. Newman, "UCI Machine Learning Repository," School of Information and Computer Science, Univ. of California, http://www.ics.uci.edu/~mlearn/MLRepository. html, 2007.

LIST OF PUBLICATIONS

1. M.Pavithra, Ms.D.Chandrakala **"**Measurement of Similarity using Link Based Cluster Approach for Categorical Data"**,** International Conference, Information Communication and Embedded Systems (ICICES'13), S.A.Engineering College Chennai, 21st February 2013.
2. M.Pavithra, Ms.D.Chandrakala "Similarity Measures for Link Based Cluster Approach for Categorical Data Clustering"**,** National Conference, Innovation on Information Technology, Bannari Amman Institute of Technology, Coimbatore, 22nd February 2013.