American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-5, Issue-12, pp-282-298
www.ajer.org
Research Paper Open Access

Measurement of Similarity Using Cluster Ensemble Approach for Categorical Data

Ms. M. Pavithra[1], Dr.R.M.S.Parvathi[2].
Assistant Professor, Department of Computer Science and Engineering, Jansons Institute of Technology, Coimbatore, India[1].
Dean PG-Studies, Sri RamaKrishna Institute of Technology, Coimbatore, India[2].

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.*

Index Terms: *Clustering, Data mining, Categorical data, Cluster Ensemble, link-based similarity, Refined matrix, C-Rank link based cluster.*

I. INTRODUCTION

DATA clustering is one of the fundamental tools we have for understanding the structure of a data set. Clustering is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters[14]. Clustering often is a first step in data analysis. Clustering is the process of discovering homogeneous groups or clusters according to a given similarity measure. Clustering maximizes intra-connectivity among patterns in the same cluster while minimizing inter-connectivity between patterns in different clusters[7]. Clustering is an important technique in discovering meaningful groups of data points. Clustering provides speed and reliability in grouping similar objects in very large datasets[3]. Most previous clustering algorithms focus on numerical data whose inherent geometric properties can be exploited naturally to define distance functions between data points. However, much of the data existed in the databases is categorical, where attribute values can’t be naturally ordered as numerical values[4]. An example of categorical attribute is shape whose values include circle, rectangle, ellipse, etc. Consensus clustering can provide benefits beyond what a single clustering algorithm can achieve. Consensus clustering algorithms often: generate better clustering’s; find a combined clustering unattainable by any single clustering algorithm. The consensus clustering algorithm can be applied to the ensemble of all clustering’s produced by discrete features of the data set[10]. Cluster ensemble (CE) is the method to combine several runs of different clustering algorithms to get a common partition of the original dataset, aiming for consolidation of results from a portfolio of individual clustering results[3]. 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[6]. 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[5]. Clustering ensembles have emerged as a powerful method for improving both the robustness as well as the stability of unsupervised classification solutions. A cluster ensemble can be defined as the process of combining multiple partitions of the dataset into a single partition, with the objective of enhancing the consensus across multiple clustering results[12]. A cluster ensemble technique is characterized by two components: the mechanism to generate diverse partitions, and the consensus function to combine the input partitions into a final clustering.

II. RELATED WORK

We briefly describe relevant work in the literature on categorical clustering and cluster ensembles. Clustering of categorical data has recently attracted the attention of many researchers[5][6]. The k-modes algorithm is an extension of k-means for categorical features. To update the modes during the clustering process, the authors used a new distance measure based on the number of mis-matches between two points. Squeezer is a categorical clustering algorithm that processes one point at the time. ROCK (Robust Clustering using links) is a hierarchical clustering algorithm for categorical data. It uses the Jaccard coefficient to compute the distance between points[7][8]. The COOLCAT algorithm is a scalable clustering algorithm that discovers clusters with minimal entropy in categorical data. COOLCAT uses categorical, rather than numerical attributes, enabling the mining of real-world datasets offered by fields such as psychology and statistics. The algorithm is based on the idea that a cluster containing similar points has an entropy smaller than a cluster of dissimilar points. Thus, COOLCAT uses entropy to define the criterion for grouping similar objects. LIMBO is a hierarchical clustering algorithm that uses the Information Bottleneck (IB) framework to define a distance measure for categorical tuples[9][10]. The concepts of evolutionary computing and genetic algorithm have also been adopted by a partitioning method for categorical data, i.e., GAClust. Cobweb is a model-based method primarily exploited for categorical data sets. Different graph models have also been investigated by the STIRR, ROCK, and CLICK techniques. In addition, several density-based algorithms have also been devised for such purpose, for instance, CACTUS, COOLCAT, and CLOPE. The Cluster-based Similarity Partitioning Algorithm (CSPA) induces a graph from a co-association matrix and clusters it using the METIS algorithm[12][13]. Hyper graph partitioning algorithm (HGPA) represents each cluster by a hyper edge in a graph where the nodes correspond to a given set of objects.

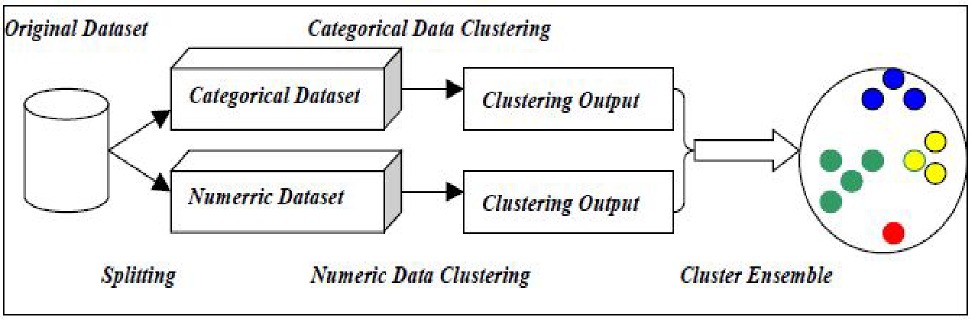
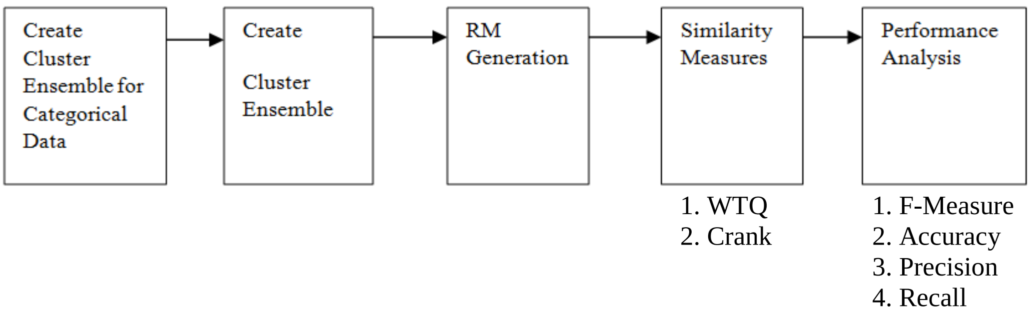


Figure 1. Overview of Cluster Ensemble Algorithm Framework.

Proposed Work

Figure.2 shows the proposed rule based similarity using link based cluster approach . A brief explanation about its phases is followed.



Module Description

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. Conventional approach, the technique developed in acquires a cluster ensemble without actually implementing any base clustering on the examined data set[14]. In fact, each attribute is considered as a base clustering that provides a unique data partition. In particular, a cluster in such attribute-specific partition contains data points that share a specific attribute value (i.e., categorical label). Thus, the ensemble size is determined by the number of categorical labels, across all data attributes. The final clustering result is generated using the graph-based consensus techniques. Specific to this so-called "direct" ensemble generation method, a given categorical data set can be represented using a binary cluster-association matrix.

2. Creating a Cluster Ensemble

Type I (Direct ensemble)

The First 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 …. x_n} be a set of N data points, A ={a_1…. a_m} be a set of categorical attributes, and $\pi = \{\pi_1 \ldots \ldots \pi_M\}$ be a set of M partitions. Each partition π_i is generated for a specific categorical attribute a_i € A. Clusters belonging to a partition πi = { C^i_1……………………C^i_{ki}} correspond to different values of the attribute $a_i = \{a^i_1, \ldots, a^i_{k_i}\}$, where $\bigcup_{j=1}^{k_i} C^i_j = a_i$ and k_i is the number of values of attribute ai. With this formalism, categorical data X can be directly transformed to a cluster ensemble, without actually implementing any base clustering[14]. While single-attribute data partitions may not be as accurate as those obtained from the clustering of all data attributes, they can bring about great diversity within an ensemble. Besides its efficiency, this ensemble generation method has the potential to lead to a high-quality clustering result.

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 number 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 $q = q_{min} + └ \alpha(q_{max} - q_{min}) ┘$. where α € [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. An attribute is selected one by one from the pool of d attributes, until the collection of q is obtained. The index of each randomly selected attribute is determined as h=└1+βd┘, given that h denotes the h^{th} attribute in the pool of d attributes and β € [0,1] is a uniform random variable[7]. Note that k-modes is exploited to create a cluster ensemble from the set of subspace attributes, using both Fixed-k and Random-k schemes for selecting the number of clusters.

III. 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. Each entry in this matrix represents a association degree between data point. 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[12][14]. For example of cluster ensemble and the corresponding BM, a large number of entries in the BM are unknown, each presented with "0." Such condition occurs when relations between different clusters of a base clustering are originally assumed to be nil. In fact, each data point can possibly associate to several clusters of any particular clustering. 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, if $c_l=C_t(xi)$,
{ $Sim(c_l,C_t(xi))$, 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, $sim(C_x,C_y)$ €[0,1] denotes the similarity between any two clusters C_x, C_y, which can be discovered using the following link-based algorithm. Note that, for any clustering π_t € π, 1<=∑

$RM(x_i,C)<=k_t$. Unlike the measure of fuzzy membership, the typical constraint of $\sum RM(x_i,C)= 1$ is not appropriate for rescaling associations within the RM.

IV. WEIGHTED TRIPLE-QUALITY (WTQ): A 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 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 . The weight assigned to the edge that connects clusters is estimated by the proportion of their overlapping[9][11]. Members Shared neighbours have been widely recognized as the basic evidence to justify the similarity among vertices in a link network. The method gives high weights to rare features and low weights to features that are common to most of the pages. For WTQ can be modified to discriminate the quality of shared triples between a pair of clusters in question[14]. The quality of each cluster is determined by the rarity of links connecting to other clusters in a network.

$$W_{xy} = \frac{|L_x \cap L_y|}{|L_x \cup L_y|}$$

ALGORITHM: $WTQ(G,C_x,C_y)$

$G = (V,W)$, a weighted graph, where $C_x,C_y \in V$;

$N_k \subseteq V$, a set of adjacent neighbors of $C_k \in V$;

$$W_k = \sum_{C_t \in N_k} W_{tk};$$

WTQ_{xy}, the WTQ measure of C_x {and} C_y;

(1) WTQ_{xy} <------------- 0

(2) For each c € N_x

(3) If c € N_y

(4) WTQ_{xy} <-------WTQ_{xy} + $(1/W_e)$

(5) Return WTQ_{xy}

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

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

V. CONNECTOR-BASED SIMILARITY MEASURE

Connector-based similarity measure called C-Rank. C-Rank uses both in-links and out-links at the same time. A new link-based similarity measure called C-Rank, which uses both in-link and out-link by disregarding the direction of references. C-Rank, where L(p) denotes the set of undirected link neighbors of paper p. Similar to that the accuracy of Co-citation (Coupling) is improved by iterative SimRank (rvs-SimRank), Crank is defined iteratively. C-Rank unifies in-links and out-links into undirected links. C-Rank has the effect similar to increasing the weight of Co-citation (SimRank) when computing the score between old papers, increasing the weight of Coupling (rvs-SimRank) when computing the score between recent papers, and increasing the weight of a BP-based similarity measure when computing the score between old and recent papers. The user does not have to set the value of C-Rank. In experiments, we show that the accuracy of C-Rank is higher than those of Amsler (P-Rank) with different values. Treating both in-links and out-links as undirected might be thought to result in loss of semantics of the direction of links. Existence and uniqueness guarantee that there exists a unique solution to iterative C-Rank which reaches a fixed point by iterative computation. C-Rank achieves a higher effectiveness than existing similarity measures in most cases. C-Rank converges at the 9-th iteration [14][15]. When C is low, the recursive power of C-Rank is weakened such that only the papers in local or near-local neighborhood are used in similarity computation. When C is high, more papers in a more global neighborhood can be used in computing the similarity recursively. When C is high, therefore, the convergence takes more time.

1: **Procedure** unweighted CRank(G, L)
2: add G to L
3: if G is a clique or a singleton return
4: S: = sparsest vertex separator of G
5: $A_1, \dots, A_k$:= connected components of G \ S

6: for i = 1 to k do
7: Gi: = sub-network of G induced on S [A_i
8: if Gi not already in L then
9: un weighted C Rank(G_i , L)

Performance Evaluation

This section presents the evaluation of the proposed link based method (LCE), using a variety of validity indices and real data sets. The quality of data partitions generated by this technique is assessed against those created by different categorical data clustering algorithms and cluster ensemble techniques.

Investigated Data Sets

The experimental evaluation is conducted over two data sets. The "UCI Machine learning repoistory" data set is a subset of the well known attribute values collection— UCI Machine learning repository.

Data Normalization

A summary of the datasets taken from the UCI Machine learning repository is shown in Table 1. The datasets are selected in such a way that the problems chosen are with at least six classes and no missing values.

Table 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 (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:

$$A_i = \frac{t}{n} \times 100 \qquad (1)$$

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 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. 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 4.

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

Table 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 Type II Type III	70.93 75.35 81.89	71.05 77.29 81.90	87.59 88.01 89.43	93.35 93.54 94.49
4	Type I Type II Type III	70.95 75.68 81.87	71.28 77.33 85.76	87.62 87.75 91.76	94.29 96.67 96.72
5	Type I Type II Type III	70.97 76.56 79.45	71.22 78.89 85.34	86.72 88.18 90.76	94.82 96.75 96.83
6	Type I Type II Type III	70.92 76.87 84.29	71.60 81.33 89.37	86.84 89.33 92.83	95.78 97.73 97.89
7	Type I Type II Type III	70.21 83.88 84.67	72.55 88.78 89.56	87.72 92.77 93.56	95.61 97.77 99.20

Table 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 Type II Type III	72.65 77.31 83.62	73.93 79.89 84.39	88.98 89.99 89.97	94.98 95.44 95.85
4	Type I Type II Type III	72.66 77.35 82.89	73.96 79.92 86.89	89.96 89.99 92.80	95.44 97.77 97.96
5	Type I Type II Type III	72.71 78.88 80.96	74.09 80.56 86.78	88.89 90.89 92.67	95.88 97.92 97.98
6	Type I Type II Type III	72.48 78.92 85.99	73.82 83.66 90.67	89.91 91.88 93.44	96.50 97.94 98.92
7	Type I Type II Type III	72.77 84.90 85.89	75.20 90.22 90.99	89.90 93.97 94.78	96.84 98.90 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.

$$\text{Precision } (i,j)= n_{ij} / n_j \qquad (2)$$

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 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 (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 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 7.

Table 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 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 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,

$$\text{Recall }(i,j) = n_{ij} / n_i \qquad (3)$$

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 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 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 10.

Table 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 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 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 Type II Type III	74.78 84.93 87.91	76.27 90.56 91.99	89.94 93.98 94.80	96.89 98.95 99.68

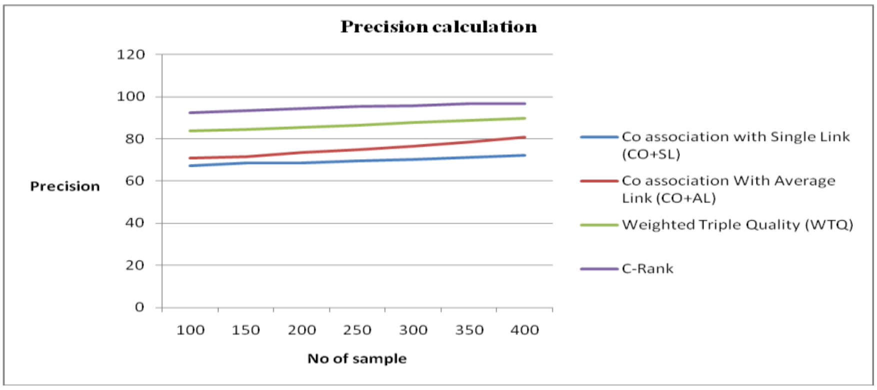


Fig.11 Graph for Performance of precision based on number of samples

The above graph in the Fig.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.

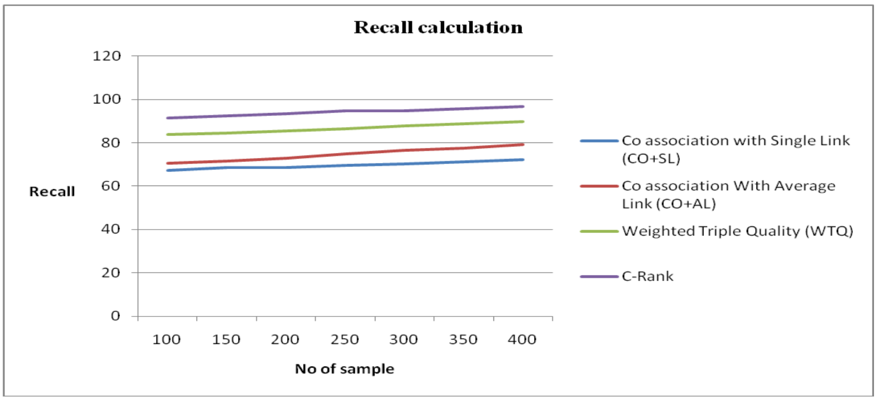


Fig.12 Graph for Performance of Recall rate based on number of samples

The above graph in the Fig.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.

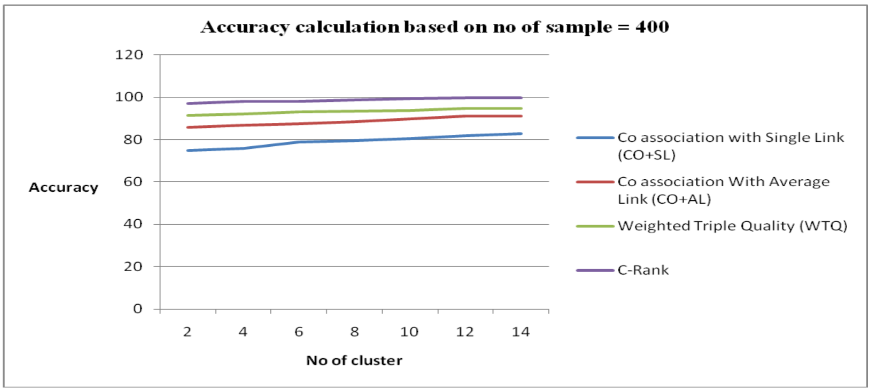


Fig.13 Graph for Performance of Accuracy based on number of sample = 400.

The above graph in the Fig.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.

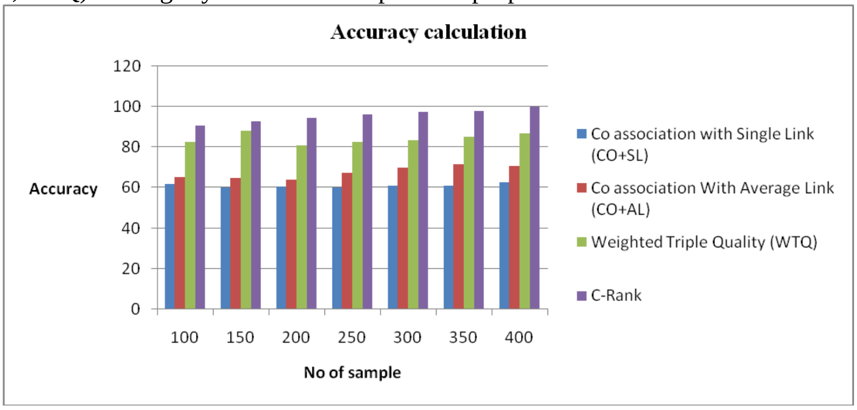


Fig.14 Graph for Performance of Accuracy based on number of samples

The above graph in the Fig.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.

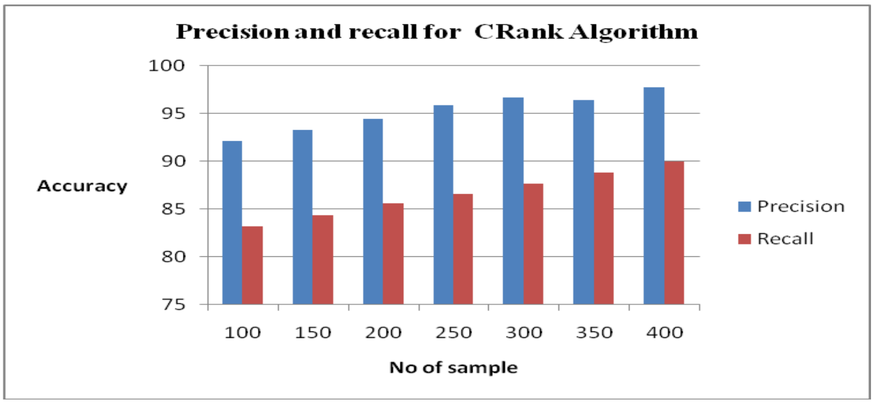


Fig.15 Graph for comparison of precision and recall for C-Rank algorithm

The above graph in the Fig.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.

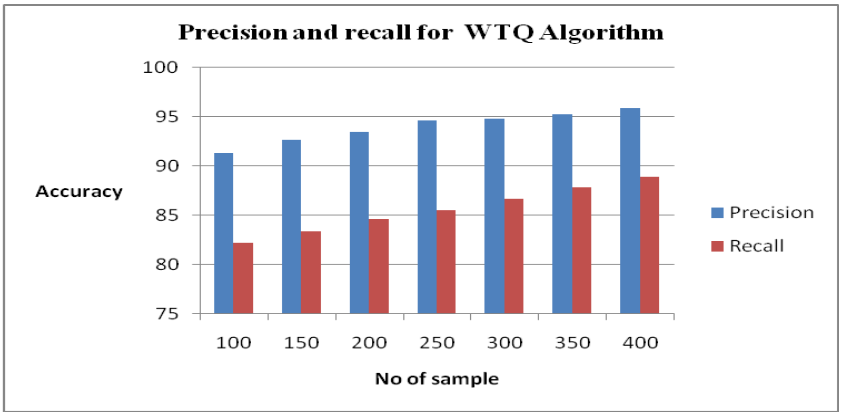


Fig.16 Graph for comparison of precision and recall for WTQ algorithm

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.16.

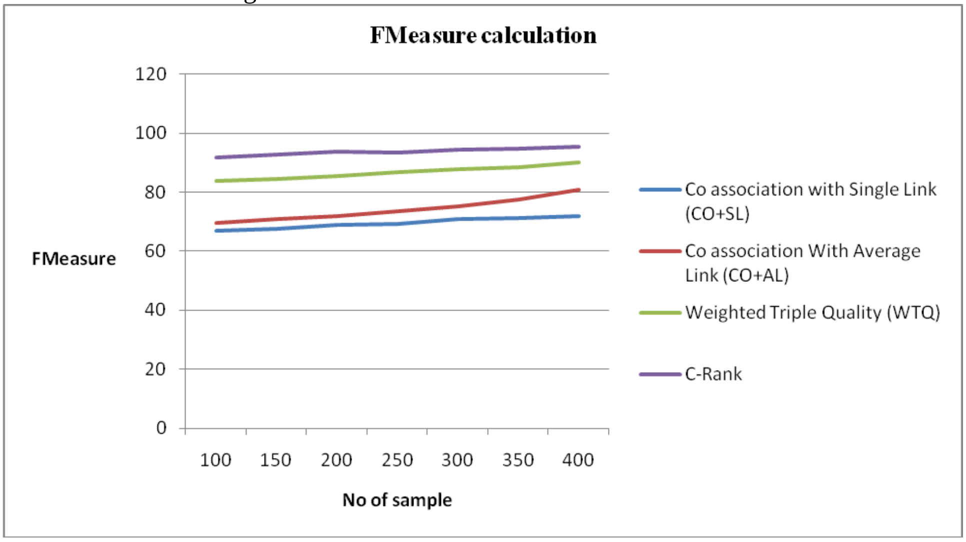


Fig.17 Graph for Performance of FMeasure based on number of samples

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.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 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. 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 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 20.

Table 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 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 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 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 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 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 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 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 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

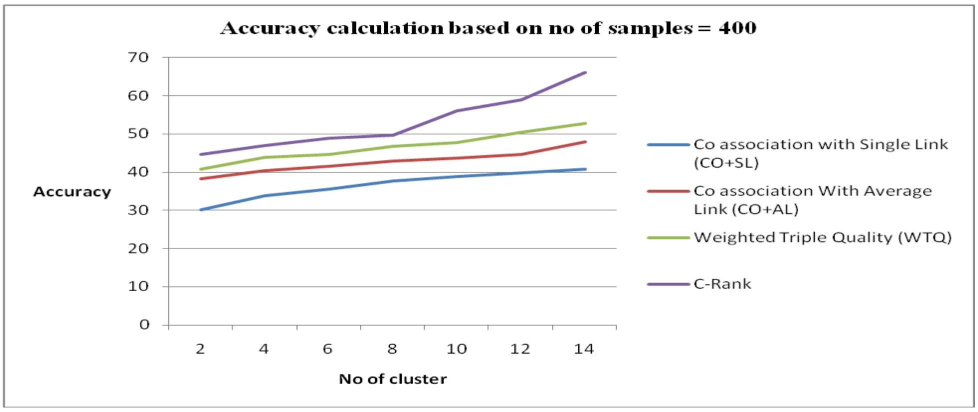


Fig.27 Graph for Performance of Accuracy based on number of sample = 400.

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.27.

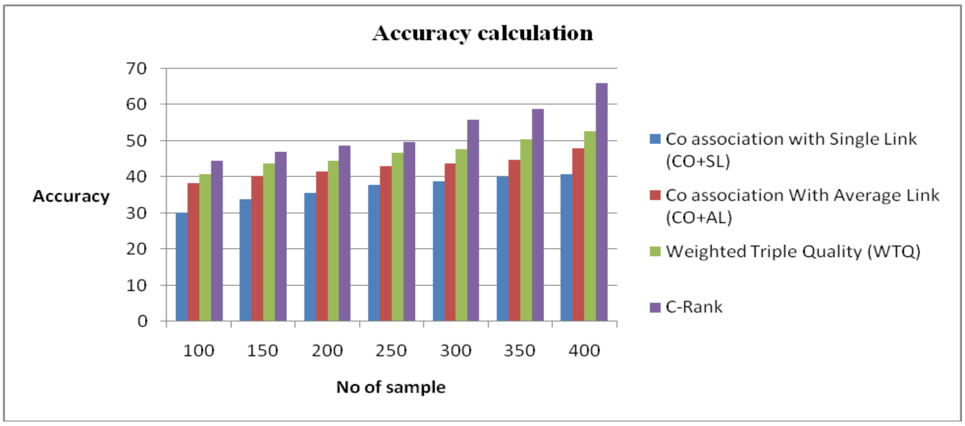


Fig.28 Graph for Performance of Accuracy based on number of samples

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.28.

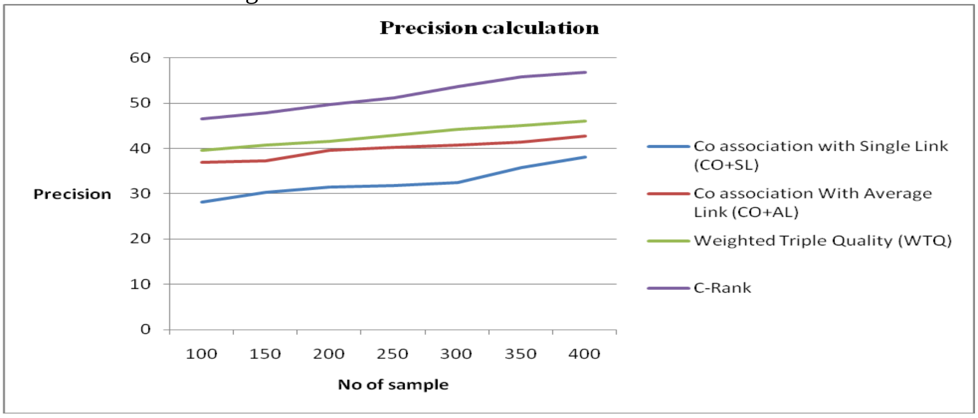


Fig.29 Graph for Performance of Precision based on number of samples

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.29.

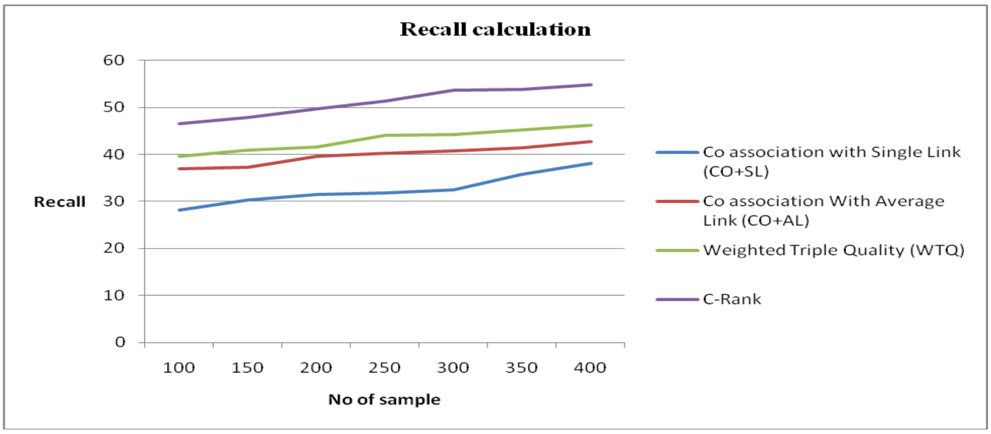


Fig.30 Graph for Performance of Recall based on number of samples

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.30.

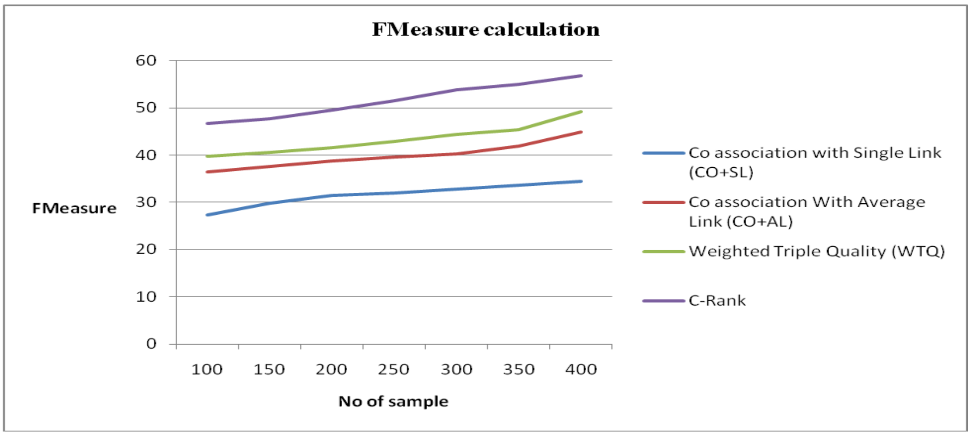


Fig.31 Graph for Performance of F Measure based on number of sample

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.31.

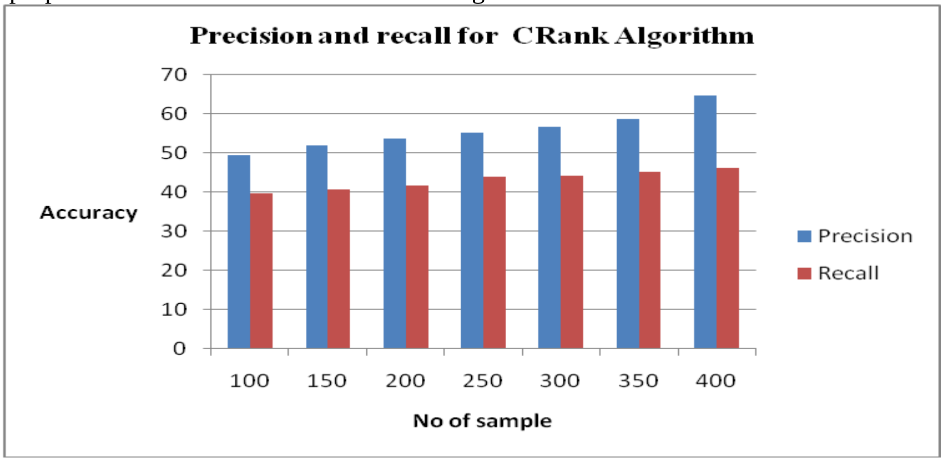


Fig.32 Graph for comparison of precision and recall for C-Rank algorithm

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.32.

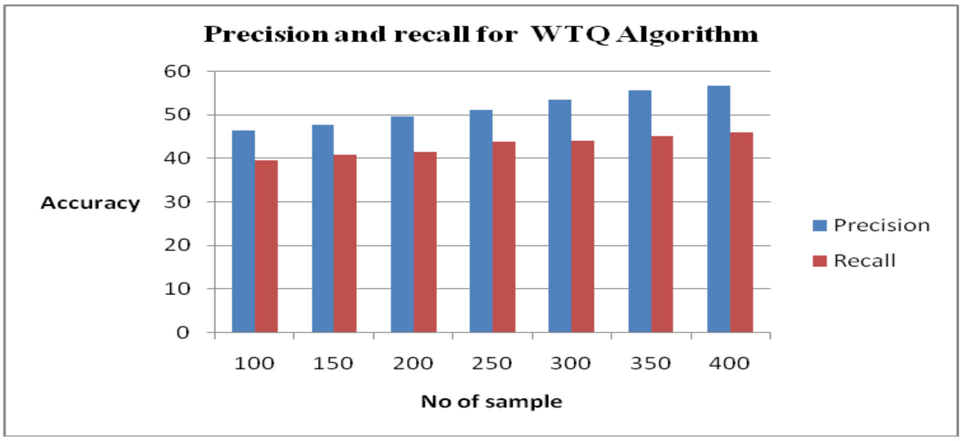


Fig.33 Graph for comparison of precision and recall for WTQ algorithm

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.33.

VI. CONCLUSION

This paper presents a novel, highly effective link-based cluster ensemble approach(WTQ) 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. It also presents a Crank link based cluster approach for categorical data clustering.

VII. 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.

REFERENCES

[1]. N. Nguyen and R. Caruana, "Consensus Clusterings," Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 607-612, 2007.

[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]. C. 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]. A. Strehl and J. Ghosh, "Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions," J. Machine Learning Research, vol. 3, pp. 583-617, 2002.

[5]. Z. He, X. Xu, and S. Deng, "A Cluster Ensemble Method for Clustering Categorical Data," Information Fusion, vol. 6, no. 2, pp. 143-151, 2005.

[6]. 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.

[7]. Y. Zhang, A. Fu, C. Cai, and P. Heng, "Clustering Categorical Data," Proc. Int'l Conf. Data Eng. (ICDE), p. 305, 2000.

[8]. M. Dutta, A.K. Mahanta, and A.K. Pujari, "QROCK: A Quick Version of the ROCK Algorithm for Clustering of Categorical Data," Pattern Recognition Letters, vol. 26, pp. 2364-2373, 2005.

[9]. E. Abdu and D. Salane, "A Spectral-Based Clustering Algorithm for Categorical Data Using Data Summaries," Proc. Workshop Data Mining using Matrices and Tensors, pp. 1-8, 2009.

[10]. B. Mirkin, "Reinterpreting the Category Utility Function," Machine Learning, vol. 45, pp. 219-228, 2001.

[11]. A.P. Topchy, A.K. Jain, and W.F. Punch, "A Mixture Model for Clustering Ensembles," Proc. SIAM Int'l Conf. Data Mining, pp. 379-390, 2004.

[12]. 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.

[13]. 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.

[14]. Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, and Chris Price," A Link-Based Cluster Ensemble Approach for Categorical Data Clustering," IEEE Transactions On Knowledge And Data Engineering, Vol. 24, NO. 3, March 2012.

[15]. A. Topchy, A. K. Jain, and W. Punch. Combining multiple weak clusterings. In Proceedings IEEE International Conference on Data Mining, pages 331–338, 2003.

[16]. C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998. URL http://www.ics.uci.edu/∼mlearn/MLRepository.html.