Weighted Delta Factor Cluster Ensemble Algorithm for Categorical Data Clustering in Data Mining

Sarumathi Sengottaiian[1], Shanthi Natesan[2], and Sharmila Mathivanan[3]
[1]Department of Information Technology, K.S.R College of Technology, India
[2]Department of Computer Science and Engineering, Nandha Engineering College, India
[3]Department of Information Technology, M.Kumarasamy College of Engineering, India

Abstract: *Though many cluster ensemble approaches came forward as a potential and dominant method for enhancing the robustness, stability and the quality of individual clustering systems, it is intensely observed that this approach in most cases generate a final data partition with deficient information. The primary ensemble information matrix generated in the traditional cluster ensemble approaches results only the cluster data point relations with unknown entries. This paper mainly denotes the improved analysis of the Link based Cluster Ensemble (LCE) approach which overcomes the problem of degrading the quality of clustering result and in particular it presents an efficient novel Weighted Delta Factor Cluster Ensemble algorithm (WDFCE) which enhances the refined matrix by augmenting the values of similitude measures between the clusters formed in the Bipartite cluster graph. Subsequently to obtain the final ultimate cluster result, the pairwise-similarity consensus method is used in which K-means clustering technique is applied over the similarity measures that are formulated from the Refined Similitude Matrix (RSM). Experimental results on few UCI datasets and synthetic dataset reveals that this proposed method always outperforms the traditional cluster ensemble techniques and individual clustering algorithms.*

Keywords: *Clustering, cluster ensembles, consensus function, data mining, refined matrix, similitude measures.*

Received October 28, 2013, accepted November 4, 2014

1. Introduction

Cluster ensembles bid a solution to the challenges and issues arising from the ill-posed behaviour of the clustering algorithms. As data clustering is one of the most essential factor and an underpinning process in data mining, it also plays an imperative role in the other fields such as: machine learning process, pattern recognition, information retrieval, spatial data extraction, image processing, networking and World Wide Web. The main objective of the cluster analysis is finding similarities between data according to the uniqueness found in the data and grouping related data objects into clusters. An excellent clustering produces a high quality clusters with maximized intra class similarity and minimized inter class similarity.

A large variety of clustering algorithms which are of well established such as K-means, Expectation Maximization (EM) based on the spectral graph theory [33], K-modes, Genetic Algorithm for Clustering (GAClust) [7], Sieving Through Iterated Relational Reinforcement (STIRR) [16], uses different graph models for clustering, RObust Clustering using linKs (ROCK) employs links for clustering [17], CLICK [39] finds clusters in categorical datasets based on k-partite maximal cliques, Clustering Categorical Data Using Summaries (CACTUS) [13], COOLCAT [4] Entropy-based algorithm for categorical clustering, CLOPE [37] clustering large transactional databases with high dimensions, Squeezer [20] uses prespecified threshold for clustering categorical data, and also clustering high dimensional data using constraint-partitioning K-means [14] clustering algorithms, differential fuzzy clustering, standard deviation of standard deviation roughness algorithm, frequency of attribute value combination algorithm and some hierarchical clustering algorithms like divisive algorithm (LIMBO) [1], single link, fuzzy C-means, fuzzy C-medoids [26] etc., are emerged over earlier periods. Conversely, it is known that there is no single clustering method is capable of providing accurate and appropriate cluster results [16]. Since by applying a clustering algorithm to the data set it works on the basis of the internal criteria i.e., similarity or dissimilarity measures used in that algorithm. Therefore, this critical concern is very difficult to evaluate the exact clustering results.

In cluster analysis the evaluation of the results are associated to the use of cluster validity indexes which is used to measure the quality of clustering results [16]. Nevertheless, to overcome this serious issue combining multiple clustering approaches in an ensemble framework may allow one to take advantage of the strengths of individual clustering approaches. The general outline of the cluster ensemble is done by achieving the solutions from the different base clustering which are then aggregated to form a final partition [26]. Some examples of the well-known cluster ensemble techniques are as follows:

- Weighted cluster ensemble [8] methods that discovers clusters in subspaces spanned by different combinations of dimensions through local weightings of features.
- Direct approach [5] that obtains the final solution through relabeling the base clustering.
- Fuzzy cluster ensemble [31] method that makes use of the relationship degree between different attributes for pruning a part of features.
- Bayesian cluster ensemble [36] method that deals with Bayes' theorem with two distinct interpretations.
- Graph based cluster ensemble methods [9] utilize a graph partitioning methodology.
- Spectral clustering ensemble [28] approach that is mainly based on resampling technique.
- Exact method based cluster ensemble [6] technique mainly recombines the partially generated solutions of different base clustering.
- Projective clustering ensemble [18] mainly deals with the subsets of input data having different subsets of features correlated to them.
- Pairwise similarity approach [3] that makes co-occurrence relations between the data objects.

In spite of the notable success these above methods generate the final data partition with deficient information of a cluster ensemble. The techniques used in the traditional ensemble information matrix deals only with the cluster data point relationships whereas it entirely ignores those among the clusters. As a result many subsisting ensemble techniques similarity matrix entries are left unknown. This paper introduces the novel Weighted Delta Factor Cluster Ensemble (WDFCE) approach which drastically improves the link based approach [26] by augmenting the similarity measures in the refined matrix. This approach along with the linked cluster network concept [26] also enhances the ability of ensemble methodology for categorical data, which has not been more popular in the past ensemble methods. Additionally, it also examined experimentally that the proposed approach is generic such that it can be efficiently applied to other data types.

The remaining part of this paper is systematized as follows: section 2 presents the general outline of the cluster ensemble methodology in which it includes generation methods and consensus functions upon which this approach has been established. Section 3 introduces the proposed Weighted Delta Factor (WDF) approach including its working paradigm and the improved Refined Similitude Matrix (RSM). Section 4 reveals the performance evaluation of this new technique compared with the traditional ensemble methods over the categorical, integer, real world datasets and also with a synthetic dataset which contains the NAC-Tech Scores of the college candidates. This paper is concluded in section 5 along with the implication for future work.

2. Cluster Ensemble Paradigm

Cluster ensembles have emerged as a recent issue of classifier ensemble exploration [8]. The fundamental idea is to combine the solutions of the various weak clustering algorithms to obtain the ultimate clustering of the dataset and it proves to be better than the individual cluster results. It also provides for a visualization tool to examine cluster number, membership and boundaries. This meta level approach involves the two major tasks of generating a cluster ensemble and then producing a final partition normally referred as the consensus function [8, 26]. Precisely the great challenge in clustering ensemble is the definition of most suitable consensus function which is capable of improving the consequences of single clustering algorithm. The basic process of the cluster ensemble is shown in Figure 1.

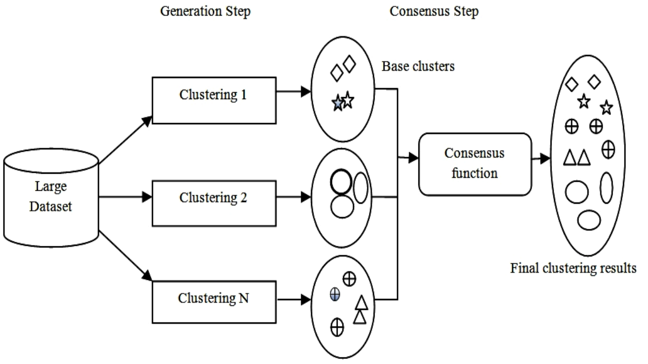


Figure 1. Basic process of cluster ensembles.

2.1. Ensemble Generation Process

Ensembles are more efficient, when assembled from a set of forecaster whose errors are dissimilar [30]. To a huge extent, miscellany among the ensemble methods will enhances the result of cluster ensemble. In particular the results obtained from clustering the dataset using any single clustering algorithm over much iteration are usually similar to each other. This circumstance leads all the ensemble members to concur with the process of partitioning the dataset.

As a result several approaches have been proposed to introduce the synthetic volatility in clustering algorithms, which paves the way for multiplicity within the cluster ensemble. The following subsequent ensemble generation methods defer different clusterings of the same data, by developing cluster models and data partitions.

- *sFixed-K*: This technique creates the fixed number of clusters (k) for each ensemble member [23].
- *Random-K*: This technique creates the randomized number of clusters (k) for each ensemble member [12].

- *Projection of Data on Different Subspace/Sampling*: The cluster ensemble can be achieved by producing base clusters from different object representations or subsets of objects [10] of the initial dataset. It can also be obtained from different subspaces, features and data sampling [35].
- *Homogeneous Ensembles*: Base clustering solutions are achieved through the repeated usage of the single clustering algorithm with various different parameter declarations such as cluster center point of the K-means algorithm [34].
- *Heterogeneous Ensembles*: Number of different clustering algorithms is used mutually to generate the base clustering results [22].
- *Mixed Heuristics*: This method [25] results in using the any combination of the aforesaid techniques to generate the base clusters.

2.2. Consensus Functions

A diverse collection of consensus functions have been developed and made accessible for extracting the final data partitions after the formation of cluster ensemble. Each consensus technique exploits the specific form of information matrix in which it précis the base clustering results. In spite of this background the consensus functions can be categorized into different forms as follows:

- *Direct Approach:* This approach is based on the relabeling and searching for the final partition that has been the best match for all ensemble members. It generates the unique set of decision labels from the heterogeneous clustering decisions.
- *Cluster based Similarity Partitioning Approach*: Here the similarity between the data objects are directly proportional to the number of ingredient clusterings of the ensemble in which they are aggregated together. The more similar data points are credited with higher chance to be placed in the same cluster. The computational and storage complexity of this method is quadratic in nature.
- *Hyper-graph Partitioning Approach*: In this approach the formulation of the cluster ensemble problem is done as partitioning the hyper-graph by dividing the minimal number of hyper edges.
- *Meta Clustering Approach*: This technique [29] initially solves the cluster correspondence problem by grouping the clusters indentified in the individual clustering solutions. After that it uses the voting method to set the data points into final consensus cluster results.
- *Pairwise Similarity Approach*: This approach [38] generates a matrix containing similitude measures among the paired data points through which any similarity based base clustering algorithm can be applied.
- *Graph-based Approach*: Graphical representation of similarity measures of the data points is created from a Pairwise matrix. To achieve the final clustering result the graph is partitioned into finite number of estimated equal sized partitions using METIS [29] or Spectral graph partitioning technique [32].
- *Feature based Approach*: It deals with the cluster label generated as an outcome of each base clustering such that it was considered as a new feature describing each data point in which it is used to originate the vital cluster solution [7].

3. A Novel WDFCE Approach

On hand Cluster Ensemble methods for clustering categorical data rely on the classical Pairwise-similarity and the Binary co-association Matrix (BM) [19] in which it reviews the underlying ensemble information at a relative coarse level. Many matrix value entries are left blank and simply filled with “0”. Due to this issue the quality of the clustering results are degraded to a large extent. To overcome this concern, a new method namely WDFCE algorithm along with the link based concept has been established to discover the unknown values, thereby enhancing the measures of refined matrix of link based approach [26] and hence in turn it improves the accuracy rates of the ultimate cluster partition.

This approach is more efficient than the traditional cluster ensemble methods as many of them mainly focus on BM matrix of similarity measures where accuracy levels are not appropriate to the great extent. So the novelty moves to estimate the similarity among the cluster partitions rather examining the data points. A new WDFCE algorithm has been purposely exposed to produce Similitude measures in an accurate and inexpensive manner.

The WDFCE methodology is ilustrated in Figure 2. It includes three major steps of:

- Generating base clustering results to form the cluster ensemble (P).
- Generating a RSM using the WDF algorithm.
- Extracting the ultimate data partition (P*) by utilizing the pair wise similarity technique as a consensus function.

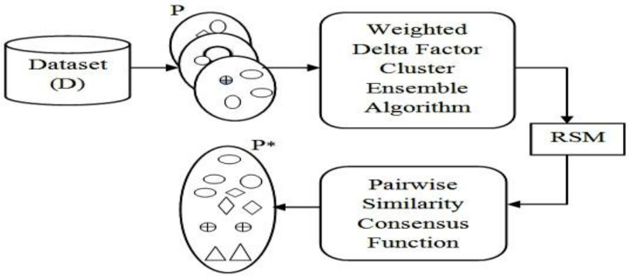

Dataset (D)
P
Weighted Delta Factor Cluster Ensemble Algorithm
RSM
P*
Pairwise Similarity Consensus Function

Figure 2. WDFCE framework.

3.1. Creating Cluster Ensemble

Let $D=\{d1, ..., dn\}$ be a set of data points and P be the cluster ensemble such that $P=\{P_1, \ldots, P_M\}$ are the ensemble members with M base clustering's. Each base clustering profits a set of clusters $Pi=\{C_1^i, C_2^i, \ldots, C_k^i\}$, where as k_i is number of clusters in the i^{th} clustering results. The following Figure 3 illustrates the cluster ensemble [26] and its corresponding clusters.

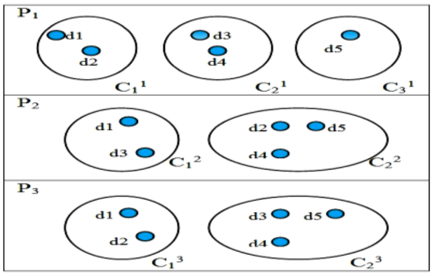
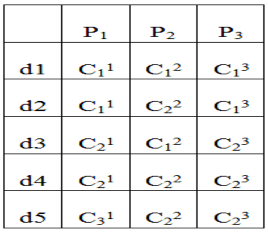


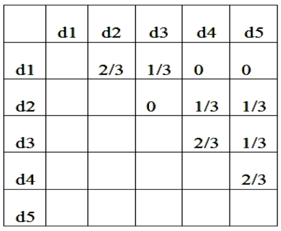
Figure 3. Sample cluster ensemble.

For this approach the K-means algorithm [21] is used to generate the base clustering results in which different randomized parameter initialization of the cluster centers are applied. In particular to a full-space ensemble, productions of base clusterings are created from the original dataset with all the attributes and the instances. In order to provide the efficient results the two schemes such as fixed-K [23] where $K=[\sqrt{D}]$ in which D is number of data points, and random-K [12, 28] where $K\in\{2, \ldots, [\sqrt{D}]\}$ (here randomized number of clusters are generated to each ensemble member) are employed to determine the number of clusters obtained in the base clustering solutions. From the sample cluster ensemble [26] shown in Figure 3, first label assignment matrix as shown in Figure 4 of size $D\times M$ was created. It mainly symbolizes the cluster labels that are assigned to each data objects by different base clusterings. Second, the Pairwise similarity matrix in Figure 5 of size $D\times D$ précis the statistics among the data objects occurred.



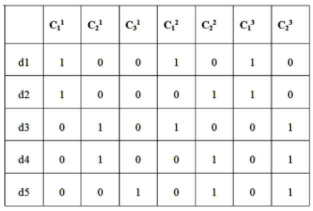
	P₁	P₂	P₃
d1	C_1^1	C_1^2	C_1^3
d2	C_1^1	C_2^2	C_1^3
d3	C_2^1	C_1^2	C_2^3
d4	C_2^1	C_2^2	C_2^3
d5	C_3^1	C_2^2	C_2^3

Figure 4. Label-assignment matrix.



	d1	d2	d3	d4	d5
d1		2/3	1/3	0	0
d2			0	1/3	1/3
d3				2/3	1/3
d4					2/3
d5					

Figure 5. Pairwise Similarity matrix.



	C_1^1	C_2^1	C_3^1	C_1^2	C_2^2	C_1^3	C_2^3
d1	1	0	0	1	0	1	0
d2	1	0	0	0	1	1	0
d3	0	1	0	1	0	0	1
d4	0	1	0	0	1	0	1
d5	0	0	1	0	1	0	1

Figure 6. Binary cluster association matrix.

Furthermore the binary cluster association matrix [19] in Figure 6 reveals the cluster specific nature of the original label assignment matrix. Each entry in this matrix mainly denotes the crisp association degree between the data points and the clusters formed in the ensemble. The relationship degree is based on the occurrence of the data points in the generated clusters. It records the matrix entry by either "1" or "0" such that if the particular data point is present then its corresponding entry will be recorded as "1" otherwise "0".

3.2. WDFCE Algorithm

As given in the Sample cluster ensemble P with a set of data points $D=\{d1, d2, \ldots, dn\}$ a Weighted Cluster Graph $WCG=(V, W)$ can be constructed, where V is the set of vertices representing the link between the clusters and W be the weighted factors between the edges of the clusters. The following Figure 7 represents the bipartite graph of linked network of clusters which was generated from sample ensemble given in Figure 3.

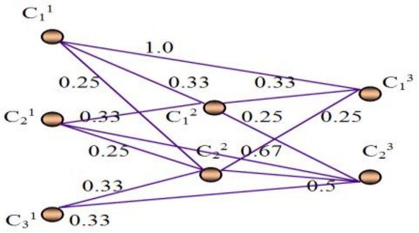


Figure 7. Sample bipartite cluster graph.

Formally the weight allotted for each cluster $W_{xy}\in W$ that connects the clusters Cx and $Cy\in V$ is calculated by the proportion of their related members in the clusters as given below:

$$W_{xy} = \frac{dx \cap dy}{dx \cup dy} \tag{1}$$

Where $dx\subset D$ denotes the set of data points corresponds to the cluster $Cx\in V$. In the graph, the circle nodes denotes the clusters and the edges present only when its appropriate weights are tends to be nonzero. Neighbours linked to each other in the cluster network recognized as a proof to express the similarity among the vertices in the network [15]. Basically the vertex $Ct \in V$ is a common neighbour also referred as the delta in which it shows the delta formation when connected as a center for the two cluster vertices. Many advanced proceedings extends this basics as an account of common neighbours that are under examination in connected path, SimRank [27] and the random walk algorithms [11].

In spite of these reported efficiency, these techniques are highly expensive and even impractical for mixed and large datasets. Henceforth, the WDF cluster ensemble algorithm is proposed mainly for the efficient similitude measures between the clusters in the linked network. Unlike other techniques, WDF aims to measure the proximity between the clusters rather than determining the data points. The superiority of each cluster is determined on the basis of its rarity of links connecting to other clusters in the linked bipartite cluster graph. WDF is mainly motivated by the relationship degree between the two clusters in contrast towards the highest accuracy rates.

With the WCG presented in Figure 7, for the clusters Cx , $Cy \in V$ the weighted factor of each cluster denotes the collection of clusters which are having direct link with that cluster, this in turn can be estimated by:

$$W_{dx} = \Sigma_{\forall Ct \in Nl} W_{tx} ; W_{dy} = \Sigma_{\forall Ct \in Nl} W_{ty} \qquad (2)$$

Where W_{dx} and W_{dy} are the linked weight factors of the clusters Cx and Cy. The WDF measure of clusters $Cx, Cy \in V$ with respect to each delta d is measured as follows:

$$WDF_{xy}^{l} = \Sigma W_{dx} + \Sigma W_{dy} \qquad (3)$$

The accrued *WDF* score from all deltas (1, …, *d*) between the clusters *Cx* and *Cy* can be evaluated as follows:

$$WDF_{xy} = \sum_{l=1}^{d} \frac{1}{WDF_{xy}^{l}} \qquad (4)$$

Here, the reciprocal of the linked weight factors are considered to calculate the Similitude measure between the clusters *Cx* and *Cy*, this can be calculated as given below in Equations 5 and 6:

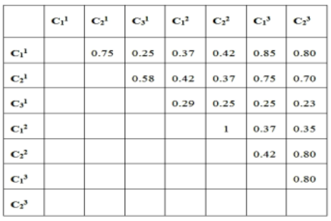
$$W_{cx} = \frac{1}{W_{dx}} ; W_{cy} = \frac{1}{W_{dy}} \qquad (5)$$

$$Sim_{WDF}(Cx, Cy) = \frac{WDF_{xy}}{(Max(|W_{cx}|, |W_{cy}|))} * DC \qquad (6)$$

```
Algorithm 1: WDF (WCG, Cx, Cy).
Input: A Dataset with d-dimensional data objects
Output: RSM
WCG = (V,W) a weighted cluster graph where Cx , Cy∈V;
N_l ⊂V, a set of adjacent neighbours of Ct∈V;
begin
$W_{dx} = \Sigma_{\forall Ct \in Nl} W_{tx}$ ; $W_{dy} = \Sigma_{\forall Ct \in Nl} W_{ty}$ ;
init $WDF_{xy}^{l} \to 0$;
  for each Ct∈Nl;
    $WDF_{xy}^{l} = \Sigma W_{dx} + \Sigma W_{dy}$ ; $WDF_{xy} = \sum_{l=1}^{d} \frac{1}{WDF_{xy}^{l}}$ ;
  end
return $WDF_{xy}$;
Compute: $Sim_{WDF}(Cx, Cy) = \frac{WDF_{xy}}{(Max(|W_{cx}|, |W_{cy}|))} * DC$ ;
end
```

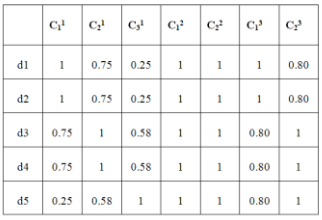
Where the similitude measures can be valued with the maximum of the reciprocated weight factor of the two clusters Cx, $Cy \in V$ and $DC \in [0, 1]$ is a stable decay value in which it is said to be the learning factor or the confidence level of recognizing the two assorted clusters being similar.

By following the sample given in Figure 3 the WDF similitude measures are estimated with the Decay Factor (DF) fixed to 0.85, and RSM was shown in Figures 8 and 9 respectively. Consequently from the emprical analysis it is proved that the reciprocled weighted values of each clusters drastically improves the refined matrix values of link based approach [15, 26].



	C_1^1	C_2^1	C_3^1	C_1^2	C_2^2	C_1^3	C_2^3
C_1^1		0.75	0.25	0.37	0.42	0.85	0.80
C_2^1			0.58	0.42	0.37	0.75	0.70
C_3^1				0.29	0.25	0.25	0.23
C_1^2					1	0.37	0.35
C_2^2						0.42	0.80
C_1^3							0.80
C_2^3							

Figure 8. WDF similarity measures between the clusters.



	C_1^1	C_2^1	C_3^1	C_1^2	C_2^2	C_1^3	C_2^3
d1	1	0.75	0.25	1	1	1	0.80
d2	1	0.75	0.25	1	1	1	0.80
d3	0.75	1	0.58	1	1	0.80	1
d4	0.75	1	0.58	1	1	0.80	1
d5	0.25	0.58	1	1	1	0.80	1

Figure 9. RSM where DC=0.85.

3.3. Applying Consensus Function to RSM

Having gained the RSM, a Pairwise Similarity method is exploited to achieve the final ultimate clustering. This Consensus technique requires the existing matrix measures to which the K-means algorithm [21] is further applied over those values to produce the final clusters among the data points. Given an RSM denoting the associations between *D* data points and clusters in the ensemble *P*, a *WCG*(*V*,*W*) can be constructed, from which by using the weighted factors of each clusters RSM measures are extracted. Finally, with these values the Pairwise similarity technique [38] is applied by considering the matrix row as a k-dimensional embedded points in order to acquire the final cluster solutions and proves to be the powerful and efficient method in attaining the nearer optimal accuracy rates.

4. Performance Evaluation

This section presents the estimation of the proposed WDFCE using few validity indices and variety of datasets. The quality of each cluster partitions acquired by this technique is evaluated against three different cluster ensemble methods.

4.1. Examined Datasets

The experimental investigation is conducted over eight datasets in which it categorized into real and synthetic datasets. Most of the datasets are taken from the UCI machine learning repository [2]. The descriptions about the datasets are as follows:

- *Breast Cancer*: This dataset is one among the three domains provided by the oncology Institute that has repeatedly listed in the machine learning literature.
- *Soybean*: This dataset includes a small subset of the original soybean database.
- *Iris*: This is perhaps the best known dataset to be found mostly in the pattern recognition. It describes about the types of iris plant in which it involves three classes, one class is linearly separable and other two are nonlinearly separable.
- *Wine*: These data are the final solutions of a chemical analysis of wines evolved in the same region in Italy but derived from three different cultivars.
- *Glass*: These data are the classification of glass types which were motivated by the criminological investigation.
- *Four Gaussian*: This dataset includes the collection of random values of two dimensional Gaussian classes.
- *Leukemia*: This dataset contains the expression levels of genes taken over limited samples. These values are similar to the colon cancer dataset.
- *Synthetic*: This synthetic dataset contains the absolute scores of the NASSCOM Assessment of Competence (NAC-Tech) test achieved by the college candidates.

The details regarding the number of instances, and attribute values of each datasets are summarized in Table 1.

Table 1. Description of datasets.

Datasets	Instances	Attributes
Four-Gaussian	800	22
Leukaemia	138	72
Glass	214	10
Soybean	307	35
Wine	178	13
Iris	150	4
Breast Cancer	683	9
NAC-Tech scores	953	16

4.2. Evaluation Criteria

The experiment set out to examine the performance of the WDFCE in contrast to some traditional cluster ensemble methods. For comparative results each cluster ensemble method splits the data partitions into number of K clusters which is then evaluated with corresponding true la bels using the following validity indices.

- *Classification Accuracy (CA)*: It is the measure of number of exactly classified data objects of the clustering results compared with the known true labels divided by the total number of data points in the datasets. This CA measures can be estimated as given below:

$$CA(\prod *) = (\sum_{i=1}^{k} (Mi)) / D \tag{7}$$

Where ∏* denotes the ultimate final partition result, *Mi* illustrates the number of data objects with the majority of the cluster label points in the cluster *i*, *D* is the total number of data objects in the dataset.

- *Rand Index*: Basically rand index is the measure of the similarity between the two data clusterings. In other words it is stated that a measure [24] of number of object pairs that exist in the same and different clusters. More formally it can also be stated as a proportional measure of the quantity of agreements and disagreements between the two partitions. It can be calculated as below:

$$RandIndex\,(R) = \frac{(a+b)}{(a+b)+(c+d)} \tag{8}$$

Where (*a*+*b*) can be considered as the number of agreements between the two clusters *Cx* and *Cy* and (*c*+*d*) can be denoted as the number of disagreements of the above stated two clusters.

4.3. Parameter Settings

To evaluate the eminence of the cluster ensemble method previously defined, they are mathematically contrasted using the following constraints of Ensemble methods displayed below:

- K-Means algorithm is specifically used for generating base cluster results
- Two types of cluster ensembles are examined in this evaluation:

1. Fixed-K.
2. Random-K.

- Ensemble size (*M*)=10 is tested.
- The quality of each method valued to a specific ensemble setting is globalized with the average of

10 runs.

- The stable DC of 0.85 is oppressed with WDF.

4.4. Compared Ensemble Methods

In order to evaluate the potential of the newly proposed WDFCE method, three traditional ensemble techniques compared are as follows,

- *Link based Cluster Ensemble* (*LCE*) [26]: Performs the discovery of the unknown values in the cluster co-association matrix using SPEC as a consensus function.
- *SimRank based Similarity Method* (*SRS*) [27]: Mainly aspires to estimate the similarity measures between the clusters based on the structural context of the adjacent neighbors using graph theoretical model.
- *Approximate SimRank based Similarity Method* (*ASRS*): Mainly scopes [40] to enhance the applicability of SRS without the process of iteration in the similarity refinement. Weighted SimRank algorithm was applied in ASRS to overcome the issues held in SimRank method.

5. Experimental Results

Based on the clustering accuracy Table 2 compares the performance of the accuracy levels of different cluster ensemble methods over several examined datasets. Hence the presented notable ensemble methods implement the fixed-K and the random-K ensemble types across the average of 10 runs.

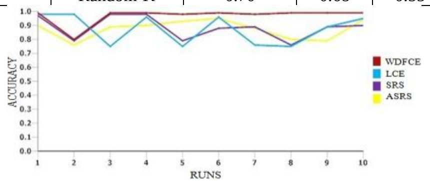
Table 2. Average CA rates of 10 runs.

Datasets	Ensemble Type	WDFCE	LCE	SRS	ASRS
Four-Gaussian	Fixed-K	0.99	0.98	0.97	0.97
	Random-K	0.98	0.94	0.96	0.95
Leukaemia	Fixed-K	0.68	0.66	0.64	0.66
	Random-K	0.66	0.64	0.63	0.61
Glass	Fixed-K	0.74	0.71	0.70	0.69
	Random-K	0.72	0.70	0.68	0.70
Soybean	Fixed-K	0.87	0.75	0.84	0.82
	Random-K	0.84	0.68	0.80	0.76
Wine	Fixed-K	0.67	0.60	0.62	0.63
	Random-K	0.79	0.67	0.68	0.70
Iris	Fixed-K	0.59	0.53	0.51	0.52
	Random-K	0.58	0.56	0.49	0.41
Breast Cancer	Fixed-K	0.56	0.50	0.50	0.49
	Random-K	0.59	0.53	0.56	0.48
NAC-Tech Scores	Fixed-K	0.73	0.69	0.68	0.62
	Random-K	0.72	0.68	0.65	0.65

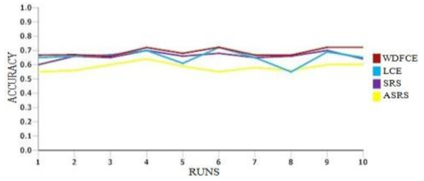
Accordingly Table 3 compares the Rand Index measures of the examined ensemble methods. The results shown in the two tables clearly explores that the WDFCE method usually performs better and produces more nearer optimal accuracy rates than the investigated cluster ensemble techniques.

Table 3. Average rand index measures of 10 runs.

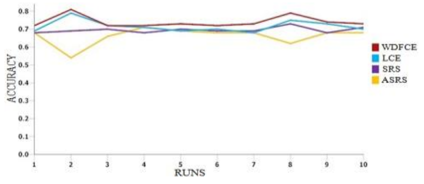
Datasets	Ensemble Type	WDFCE	LCE	SRS	ASRS
Four-Gaussian	Fixed-K	0.99	0.87	0.90	0.89
	Random-K	0.98	0.89	0.85	0.84
Leukaemia	Fixed-K	0.55	0.54	0.51	0.52
	Random-K	0.58	0.51	0.50	0.51
Glass	Fixed-K	0.58	0.57	0.40	0.49
	Random-K	0.51	0.49	0.48	0.40
Soybean	Fixed-K	0.77	0.73	0.71	0.70
	Random-K	0.71	0.70	0.63	0.68
Wine	Fixed-K	0.57	0.55	0.49	0.48
	Random-K	0.68	0.62	0.60	0.57
Iris	Fixed-K	0.66	0.63	0.65	0.64
	Random-K	0.65	0.60	0.59	0.53
Breast Cancer	Fixed-K	0.50	0.49	0.49	0.49
	Random-K	0.48	0.43	0.43	0.42
NAC-Tech Scores	Fixed-K	0.69	0.65	0.57	0.48
	Random-K	0.70	0.68	0.55	0.50



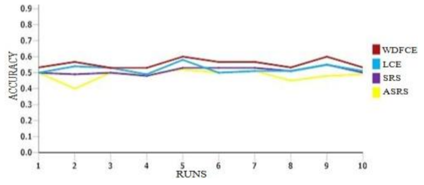
a) Four-Gaussian.



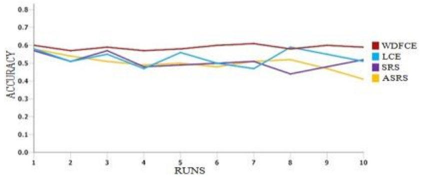
b) Leukemia.



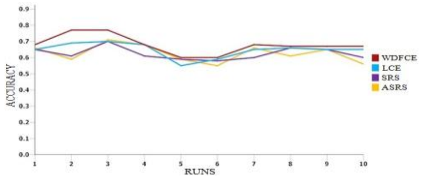
c) Glass.



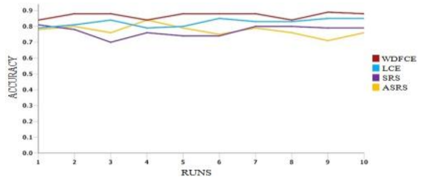
d) Breast cancer.



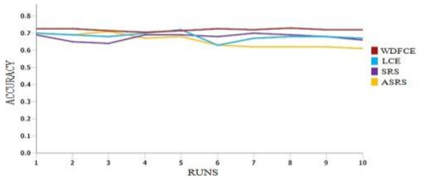
e) Iris.



f) Wine.



g) Soybean.



h) NAC-tech scores.

Figure 10. Performance of different cluster ensembles over individual datasets in accordance to accuracy rates of each runs.

Figure 10 represents the graphical expression of the performance of accuracy rates of compared cluster ensemble methods with the ensemble size M=10. And the graph indicates the accuracy rates obtained in each run of individual ensemble algorithm over the several datasets.

6. Conclusions and Future Work

Our main contribution in this paper is to exemplify the novel highly efficient WDFCE approach for categorical data clustering. It greatly transforms the original categorical data points to the numerical discrepancy of RSM, to which an effective Pairwise similarity consensus technique is directly applied to extract the final partition. The challenging issue of generating the RSM is proficiently resolved by the WDF similarity algorithm. The experiential study, of different ensemble methods with different ensemble types, validity constraints, and datasets suggests that the proposed WDF approach usually attains the superiority over the traditional standard cluster ensemble methods. The high-flying future work includes the extension of the WDFCE in neural network clustering. Furthermore this new method will also be applied to medical and biological dataset with large dimensions.

References

[1] Andritsos P. and Tzerpos V., "Information Theoretic Software Clustering," *IEEE Transactions on Software Engineering*, vol. 31, no. 2, pp. 150-165, 2005.

[2] Asuncion A. and Newman D., "UCI Machine Learning Repository," *School of Information and Computer Science, University of California, http://www.ics.uci.edu/~mlearn/MLRepository.html*, 2007.

[3] Ayad H. and Kamel M., "Finding Natural Clusters Using Multi cluster Combiner Based on Shared Nearest Neighbours," *in Proceeding of International Workshop Multiple Classifier Systems*, Guildford, pp. 166-175, 2003.

[4] Barbara D., Li Y., and Couto J., "COOLCAT: An Entropy-Based Algorithm for Categorical Clustering," *in Proceeding of The 11th International Conference on Information And Knowledge Management*, Virginia, pp. 582-589, 2002.

[5] Boulis C. and Ostendorf M., "Combining Multiple Clustering Systems," *in Proceeding of European Conference on Principles and Practice of Knowledge Discovery in Databases*, Pisa, pp. 63-74, 2004.

[6] Christou L., "Coordination of Cluster Ensembles via Exact Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 279-293, 2010.

[7] Cristofor D. and Simovici D., "Finding Median Partitions Using Information Theoretical Based Genetic Algorithms," *Journal of Universal Computer Science*, vol. 8, no. 2, pp. 153-172, 2002.

[8] Domeniconi C. and Al-Razgan M., "Weighted Cluster Ensembles: Methods and Analysis," *ACM Transaction on Knowledge Discovery Data*, vol. 2, no. 4, pp. 1-40, 2009**.**

[9] Fern X. and Brodley C., "Solving Cluster Ensemble Problems by Bipartite Graph Partitioning," *in Proceeding of International Conference on Machine Learning*, Banff, pp. 36-43, 2004.

[10] Fern X. and Brodley C., "Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach," *in Proceeding of International Conference on Machine Learning*, Washington, pp. 186-193, 2003.

[11] Fouss F., Pirotte A., Renders J., and Saerens M., "Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 355-369, 2007.

[12] Fred A. and Jain A., "Combining Multiple Clustering Using Evidence Accumulation," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 835-850, 2005.

[13] Ganti V., Gehrke J., and Ramakrishnan R., "CACTUS: Clustering Categorical Data Using Summaries," *in Proceeding of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, California, pp. 73-83, 1999.

[14] George A., "Efficient High Dimension Data Clustering using Constraint-Partitioning K-Means Algorithm," *The International Arab Journal of Information Technology*, vol. 10, no. 5, pp. 467-476, 2013.

[15] Getoor L. and Diehl C., "Link Mining: A Survey," *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 3-12, 2005.

[16] Gibson D., Kleinberg J., and Raghavan P., "Clustering Categorical Data: An Approach Based on Dynamical Systems," *Very Large Data Base Endowment Journal*, vol. 8, no. 3-4, pp. 222-236, 2000.

[17] Guha S., Rastogi R., and Shim K., "ROCK: A Robust Clustering Algorithm for Categorical Attributes," *Information Systems*, vol. 25, no. 5, pp. 345-366, 2000.

[18] Gullo F., Domeniconi C., and Tagarelli A., "Projective Clustering Ensembles," *Data Mining*

and Knowledge Discovery, vol. 26, no. 3, pp. 452-511, 2009.

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

[20] He Z., Xu X., and Deng S., "Squeezer: An Efficient Algorithm for Clustering Categorical Data," *Computer Science and Technology*, vol. 17, no. 5, pp. 611-624, 2002.

[21] Hochbaum D. and Shmoys D., "A Best Possible Heuristic for the K-Center Problem," *Math of Operational Research*, vol. 10, no. 2, pp. 180-184, 1985.

[22] Hu X. and Yoo I., "Cluster Ensemble and Its Applications in Gene Expression Analysis," *in Proceeding of Asia-Pacific Bioinformatics Conference*, New Zealand, pp. 297-302, 2004.

[23] Huang Z., "Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283-304, 1998.

[24] Hubert L. and Arabie P., "Comparing Partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193-218, 1985.

[25] Iam-On N., Boongoen T., and Garrett S., "Refining Pairwise Similarity Matrix for Cluster Ensemble Problem with Cluster Relations," *in Proceeding of International Conference on Discovery Science*, Budapest, pp. 222-233, 2008.

[26] Iam-On N., Boongeon T., Garrett S., and Price C., "A Link Based Cluster Ensemble Approach for Categorical Data Clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 3, pp. 413-425, 2012.

[27] Jeh G. and Widom J, "Simrank: A Measure of Structural-Context Similarity," *in Proceeding of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Alberta, pp. 538-543, 2002.

[28] Jia J., Xiao X., and Liu B., "Similarity-based Spectral Clustering Ensemble Selection," *in Proceeding of International Conference on Fuzzy Systems and Knowledge Discovery*, Sichuan, pp. 1071-1074, 2012.

[29] Karypis G. and Kumar V., "Multilevel K-Way Partitioning Scheme for Irregular Graphs," *Journal Parallel Distributed Computing*, vol. 48, no. 1, pp. 96-129, 1998.

[30] Kittler J., Hatef M., Duin R., and Matas J., "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239, 1998.

[31] Li T. and Chen Y., "Fuzzy Clustering Ensemble Algorithm for Partitioning Categorical Data," *in Proceeding of International Conference on Business Intelligence and Financial Engineering IEEE Computer Society*, pp. 170-174, 2009.

[32] Ng A., Jordan M., and Weiss Y., "On Spectral Clustering: Analysis and an Algorithm," *in Proceeding of Advances in Neural Information Processing Systems*, British Columbia, pp. 849-856, 2001.

[33] Vega-Pons S. and Ruiz-Shulcloper J., "A Survey of Clustering Ensemble Algorithms," *International Journal of Pattern Recognition and Artificial Intelligence* vol. 25, no. 3, pp. 337-372, 2011.

[34] Strehl A. and Ghosh J., "Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions," *The Journal of Machine Learning Research*, vol. 3, pp. 583-617, 2003.

[35] Topchy A., Jain A., and Punch W., "Clustering Ensembles: Models of Consensus and Weak Partitions," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1866-1881, 2005.

[36] Wang H., Shan H., and Banerjee A., "Bayesian Cluster Ensembles," *Statistical Analysis and Data Mining*, vol. 4, no. 1, pp. 54-70, 2011.

[37] Yang Y., Guan X., and You J., "CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data," *in Proceeding of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Alberta, pp. 682-687, 2002.

[38] Yu Z., Wong H., and Wang H., "Graph-Based Consensus Clustering for Class Discovery from Gene Expression Data," *Bioinformatics*, vol. 23, no. 21, pp. 2888-2896, 2007.

[39] Zaki M. and Peters M., "Clicks: Mining Subspace Clusters in Categorical Data via Kpartite Maximal Cliques" *in Proceeding of International Conference on Data Engineering*, Tokoyo, pp. 355-356, 2005.

[40] Zheng W., Zou L., Feng Y., Chen L., Zhao D., "Efficient Simrank Based Similarity Join Over Large Graphs," *Proceedings of the VLDB Endowment*, vol. 6, no.7, pp. 493-504, 2013.

Sarumathi Sengottaian received BE degree in Electronics and Communication Engineering from Madras University, Madras, Tamil Nadu India in 1994 and the ME degree in Computer Science and Engineering from K.S.Rangasamy College of Technology, Namakkal, Tamil Nadu, India in 2007. She is doing her PhD programme under the area Data Mining in Anna University, Chennai. She has a teaching experience of about 16 years. At present she is working as Associate professor in Information Technology department at K.S.Rangasamy College of technology. She has published 7 papers in the reputed International Journals and 2 papers in the reputed National journals. And also she has presented papers in three International conferences and four national Conferences. She has received many cash awards for producing cent percent results in university examination. She is a life member of ISTE.

Shanthi Natesan received BE degree in Computer Science and Engineering from Bharathiyar University, Coimbatore, Tamil Nadu, India in 1994 and ME degree in Computer Science and Engineering from Government College of Technology, Coimbatore, Tamil Nadu, India in 2001. She has completed PhD degree in Periyar University, Salem in offline handwritten Tamil Character recognition. She worked as a HOD in Department of Information Technology, at K.S.Rangasamy College of Technology, Tamil Nadu, India since 1994 to 2013, and currently working as a Professor and Dean in the Department of Computer Science and Engineering at Nandha Engineering College Erode. She has published 39 papers in the reputed International journals and 9 papers in the National and International conferences. She has published 2 books. She is supervising 14 research scholars under Anna University, Chennai. She acts as the reviewer for 4 International Journals. Her current research interest includes document analysis, optical character recognition, and pattern recognition and network security. She is a life member of ISTE.

Sharmila Mathivanan received BTech degree and MTech degree in Information Technology from K.S.Rangasamy College of Technology, affiliated to Anna University Chennai, Tamil Nadu, India in 2012 and 2014 respectively. At present she is working as an Assistant Professor in Information Technology Department at M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India. She has published 6 international journals and presented three papers in National level technical symposium. She is an active member of ISTE. Her Research interests include Mining Medical data, Opinion Mining and Web mining. Most of her current work involves the development of efficient cluster ensemble algorithms for extracting accurate clusters in large dimensional database.