Stochastic Modelling and Computational Sciences
Vol. 1 No. 2 (December, 2021)

Received:02nd August 2021 *Revised:10th October 2021* *Accepted: 30th October 2021*

Computational Practice for Optimal Clustering of data through Data Mining Algorithms

A.M. Golam, Department of Statistics, University of Florida, USA

Abstract

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters. However, finding a consensus clustering from multiple partitions is a difficult problem that can be approached from graph-based, combinatorial or statistical perspective. Conventional clustering algorithms utilize a single criterion to form clusters and such algorithms may not always confirm to the diverse shapes of the underlying clusters. Hence, it is necessary to perform clustering using a combined approach that combines more than one clustering technique in order to provide optimal clusters. In this paper, we propose a new clustering approach that uses the entropy values for feature selection and to apply weak clustering algorithms. We use the majority voting scheme to obtain optimal clusters from the clusters obtained from the various clustering techniques. The main advantage of our work is to enhance the clustering capability of weak clustering algorithms using the summation of cluster efficiencies in order to reduce the misclassification rate.

1. INTRODUCTION

Data mining allows extracting diamonds of knowledge from historical data and predicting outcomes of future situations. Clustering methods are often used to organize unlabelled data samples into clusters, such that the similarity among samples within a cluster and the dissimilarity among samples belonging to different clusters are maximized [9]. In this paper, we propose an optimal clustering algorithm that uses a combination of multiple partitions obtained from different clustering algorithms which improves the overall clustering efficiency for the **iris** real data set. Numerous clustering algorithms [9] proposed by various researchers are capable

Key Words: Clustering, unsupervised learning, feature selection, majority voting scheme, conflicting data points (CDP), unassigned data points (UDP), misclassification (MC) and cluster efficiency.

of producing different partitions of the same data that can capture various distinct aspects of the data set. However, a related problem is that virtually all existing clustering algorithms assume a homogeneous clustering criterion [10] over the entire feature space. As a result, all the clusters detected tend to be similar in shape and often have similar data density. Therefore, a single clustering algorithm will not find all the clusters as it utilizes a single criterion that may not confirm to the diverse shapes of the underlying clusters [2] [4] [5] [13]. The exploratory nature of clustering tasks demand efficient methods for clustering so that it would benefit from the combined strengths of many individual clustering algorithms. The major difficulty in the combined approach is finding a target partition from the output partitions from of various clustering algorithms [10]. In our work, we use an unsupervised learning method to mine unlabelled patterns so that it can be used for all types of data sets.

The main goal of our work is to find optimal clusters from **iris** data set in order to increase the classification accuracy. This work has been implemented using java and we carried out feature selection, followed by classification using three data mining algorithms namely K-means clustering algorithm, Single link algorithm and Complete link algorithm. Finally, we applied the majority voting algorithm to obtain optimal clusters. From our experimental works we found that the classification accuracy obtained from this entropy based combined approach showed less number of misclassifications in comparison with each of the three classification algorithms applied independently. The classification accuracy of our optimal clustering algorithm is about 95% which is much higher than the accuracy obtained from the clustering algorithms applied individually.

The remainder of this paper is organized as follows:

Section 2 provides a survey of related works and compares them with our work. Section 3 depicts the architecture of the system discussed in this paper and explains the various modules of the system. Section 4 shows the results obtained from our optimal clustering algorithm and compares them with other clustering algorithms. Section 5 presents the conclusions on our work and suggests some possible future enhancements.

2. RELATED WORK

Clustering is a process of grouping data items based on a measure of similarity. Clustering is a subjective process because the same set of data items often needs to be partitioned differently for different applications. This subjectivity makes the

process of clustering difficult. This is because a single algorithm or approach is not adequate to solve every clustering problem. A possible solution lies in reflecting this subjectivity in the form of knowledge [9]. In [8] [16], they describe the effectiveness of the different validity indices and clustering methods in automatically evolving the appropriate number of clusters which is demonstrated experimentally for both artificial and real-life data sets with the number of clusters varying from two to ten. In this regard, the performance of three crisp clustering algorithms, namely, hard K-Means, single linkage, and a Simulated Annealing (SA) based clustering algorithm with probabilistic redistribution of data points are also discussed. Once the appropriate number of clusters is determined, the SA-based clustering technique is used for proper partitioning of the data into the said number of clusters. Comparing with this work, we prefer a majority voting scheme than probabilistic approach. Ana L.N.Fred and Anil K.Jain [6] have addressed the problem of robust clustering based on the combination of data partitions. Using this technique, they defined objective functions and optimality criteria, based on the concept of mutual information, and they carry out variance analysis using bootstrapping. The evidence accumulation technique described by them, leads to a mapping of the clustering ensemble into a new similarity measure between patterns, by a voting mechanism on pair wise pattern associations. In their work they focused only on a single clustering algorithm which is not sufficient for effective clustering in many applications. However, we use three different clustering algorithms in order to arrive at optimal clusters.

In [2], the authors address the problem of finding consistent clusters in data partitions by proposing techniques for the analysis of the most common associations performed in a majority voting scheme. In their work, the combination of clustering results are performed by transforming data partitions into a co-association sample matrix, which maps coherent associations. This matrix is then used to extract the underlying consistent clusters. We used the majority voting scheme with the results obtained from three different clustering algorithms to obtain the combined efficiency.

In many works [1] [3] [4] [5] [10] [14], the idea of evidence accumulation for combing the results of multiple clustering has been explored. Most of the authors used K-means algorithm to decompose d-dimensional data space into large number of compact clusters to provide several clusters obtained by N random initializations and they used a knowledge framework for combining multiple partitions. In our work, we propose three different weak clustering algorithms namely K-Means, Single link and Complete link for effective clustering.

In [7] [11], feature selection has been used to reduce descriptor size in order to improve performance with respect to classification accuracy of a single clustering algorithm. However, it can be enhanced further by combining more than one clustering algorithm to get additional clustering efficiency.

In [12] [13] [15], they discussed the advantages of clustering ensembles which is a powerful method for improving both the robustness and the stability of unsupervised classification solutions. In their system, ensemble is modeled as a mixture of multivariate multi normal distributions in the space of cluster labels. However, finding a consensus clustering from multiple partitions is a difficult problem that can be solved using graph-theoretic, combinatorial or statistical approaches. Hence, we used the majority voting scheme to arrive at a consensus solution.

Comparing with works present in the literature, our work is different in many ways. First, we provide a feature selection phase to find two types of gains and an entropy calculation to obtain the best feature subset. This reduces the size of data sets and hence increases the performance in processing. Second, we apply multiple clustering algorithms to obtain initial clusters. Finally, we used the majority voting scheme to obtain the optimal set of clusters. The major contributions of our work are the combination of feature selection, clustering algorithms and majority voting into a single frame work for obtaining optimal clusters.

3. SYSTEM ARCHITECTURE

The Figure 1 shows the architecture of the system that used to obtain the optimal clusters by entropy computational model. The system consists of four modules namely feature selection module, clustering algorithm module, target clusters module and optimal cluster module.

3.1. Feature Selection

To find the best subset of attributes/features from the **iris** data set, we used C4.5 algorithm [7]. This process is shown in Figure-2.

The steps of the C4.5 algorithm is given below:

Input : Set of attributes.

Output: Subset of attributes.

Processing:

1. C4.5(instances x, attributes a)

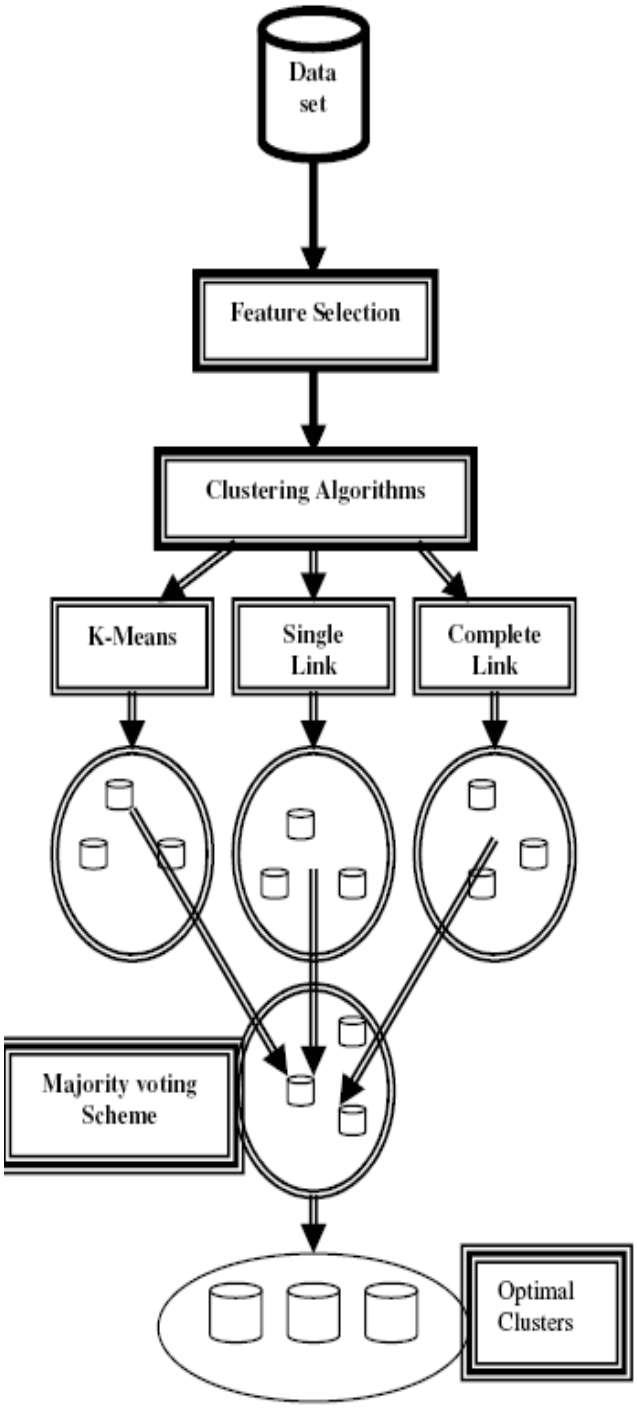

Figure 1: System Architecture

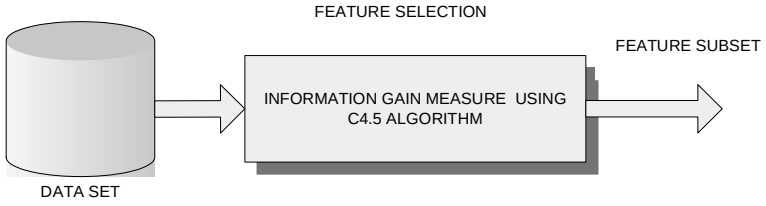

Figure 2

2. for each attribute i in a
3. if ai is continous
4. Sort x
5. Compute threshold value t = (xi+(xi+1))/2), where i = 1,2,…..n-1
6. for each threshold value Tj of ai, where j = 1,2…..n-1
7. calculate information gain Gij
8. if Gij is best gain(using entropy method) best gain = Gij and best feature = i
12. endif
13. endfor
14. endif

end for.

If one allows only binary splits, then every threshold provides unique subsets K1 and K2 of the examples at node K. The ability to choose the threshold t to maximize the splitting criterion favors continuous attributes with many distinct values. Let C denote the number of classes and p(K,j) is the proportion of cases at node K which belong to the j th class. The information at node K is computed using the relation

$$Info(K) = -\sum_{j=1}^{c}(K, j) * \log_2(p\ (K, j))$$

The information gain is calculated by a test T with L outcomes (L=2 for binary splits of continuous attributes) using the following formula.

$$\text{Gain } (K,T) = \text{Info}(K) - \Sigma^{L} \mid K_i / K \mid * \text{info } (K_i)$$

Finally the best gain is computed using distance based entropy measure. The goal of this method is to assign low entropy to intra and inter-cluster distances, and to assign higher entropy to noisy distances. The value of entropy E is computed using the formula

$$E = - Xi - Xj\ [Dij \log Dij + (1-Dij) \log (1-Dij)]$$

Where Dij is the distance between instances Xi and Xj. If the entropy value between the two distances of intra and inter clusters is low, then feature subset is considered as the best feature.

3.2. Clustering Algorithms

Clustering algorithms are capable of discovering useful but unknown classes of items from large data sets using unsupervised learning [9]. This paper proposes the use of combined approach for optimal clustering using three different clustering algorithms, with each algorithm differing in its objectives. For example, for a given data set, K-means can easily identify spherical cluster, Single Link can identify ellipsoidal clusters and Complete Link can be used for finding maximum likelihood estimate of clusters. The results obtained from these algorithms are given to the target cluster module where the majority voting algorithm is applied to find the optimal clusters.

3.2.1 Target Cluster

By using the partitions of the data produced by these clustering techniques, we form the optimal set of clusters by applying the voting algorithm. In this algorithm each pair of samples are voted for finding the associations in each independent run. The results of the clustering methods are thus mapped into an intermediate space called a co-association matrix, where each cell (i, j) represents the number of times the given sample pair has co-occurred in a cluster. In majority voting, we compare the normalized votes with a fixed threshold T in order to devise the underlying data partition and to join all such clusters. The majority voting algorithm proposed in our work is follows:

Input: N samples of K dimensional clustering ensembles.

Output: Target clusters TC.
Steps:

1. Set the co-association matrix M of order N x N as a null matrix.
2. // Produce data partitions to update the above matrix

2a. For i=1 to K do begin
Run each clustering algorithm to produce a data partition P.
Update the co-association matrix using P ie., For each sample pair (i,j) in the same cluster in partition set P M(i,j) = M(i,j)+P(i,j)

3. If any M(i,j) > T, then add P(i,j) to TC(i).

3.3. Optimal Clustering Module

To optimize the target cluster the following steps has been carried out.

1. Initialize the variables CDP,UDP and MC with zero.
2. Compute the total number of CDP, UDP and MC for each clustering algorithm.
3. Compute the sum of the number of CDP,UDP and MC for all the three clustering algorithms.
4. If total number of CDP, UDP and MC for the combined data clustering algorithm is less than the total number of CDP,UDP and MC for any clustering algorithm, then target partition has optimal clusters.

4. EXPERIMENTS AND RESULTS

In this work we proposed a new architecture to obtain optimized clusters using different existing clustering algorithms namely K-Means, Single link, and complete link. The K-Means algorithm minimizes the total within-cluster variance and tends to find spherical clusters. Single Link clustering, being based on minimum spanning tree, can find chained clusters. We used iris real dataset as input to all these three algorithms and each one has been partition into number of sub clusters with respect to random and specific clusters with user's options. We get the target clusters using majority voting scheme and finally target clusters are evaluated by conflicting data points and unassigned data points and we identified the misclassifications. Our work also minimizes the number of the conflicting and unassigned data points in target clusters. Figure 3 shows the misclassification analysis performed by three

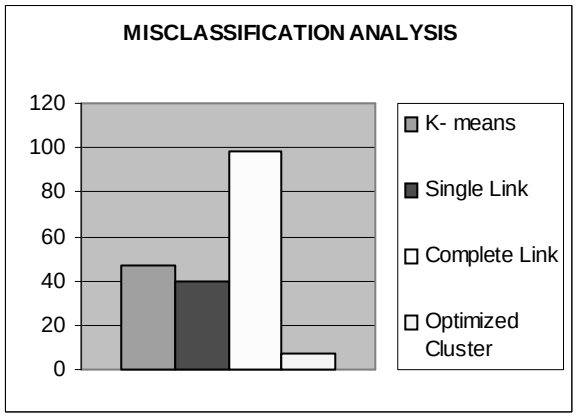

Figure 3

different clustering algorithms and also our optimal clustering algorithm. From this graph, we see that the misclassification rates are reduced in our algorithm.

5. CONCLUSIONS AND FUTURE WORKS

This paper has addressed the problem of consistent and stable clustering based on the combination of data partitions. Taking several partitions produced by various clustering algorithms, integration is designed based on the majority-voting scheme proposed in our work. The main advantages of our algorithm is that it minimizes in overcoming the conflicting points and unassigned data points. Further works carried out by us in this area are finding a suitable feature selection techniques and application of this algorithm for classification of intrusions in wired and wireless networks.

REFERENCES

[1] Alexander Strehl and Joydeep Ghosh, "Cluster Ensembles-A Knowledge Reuse Framework for Combining Multiple Partitions", *Journal of Machine Learning Research*, **3**, 2002, pp. 583-617.

[2] Ana L.N. Fred, "Finding Consisting Clusters in Data Partitions", Springer-Verlag, 2001, pp. 319-328.

[3] Ana L.N. Fred, and Jain A.K., "Combining Multiple Clusterings Using Evidence Accumulation" IEEE Transactions on Pattern Analysis and Machine Intelligence, **27**, No. 6, June 2005, pp. 835-850.

[4] Ana L.N. Fred and Jain A.K., "Data Clustering Using Evidence Accumulation", Proceedings of IEEE International Conference on Pattern Recognition-ICPR, August-2002, pp. 276-280.

[5] Ana L.N.Fred and Jain A.K., "Evidence Accumulation Clustering based on the K-Means Algorithm", Proceedings of Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops SSPR 2002 and SPR 2002, Windsor, Ontario, Canada, **2396** August 6-9, Springer-Verlag-2002, pp. 442-451.

[6] Ana L.N. Fred and Jain A.K., "Robust Data Clustering" Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, **2**, 18–20 June 2003, pp. 128-133.

[7] Devaney. M, and Ram. A, "Eeffient feature selection in conceptual clustering", Proceedings of the Fourteenth International Conference on Machine Learning, 1997, pp. 92-97.

[8] Halkidi M., Batistakis Y., Vazirgiannis M., "Clustering algorithms and validity measures". Tutorial paper, *Proceedings of SSDBM Conference*, Virginia, USA, July 2001.

[9] Jain A.K., Murthy M.N., and Flynn P., "Data Clustering" ACM Computing Surveys, **31**, No. 3, September 1999.

[10] Law M., Topchy A., Jain A.K., "Multiobjective Data Clustering, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **2**, 2004, pp. 424-430.

[11] Martin H. C. Law, Mário A., Figueiredo T., Jain A.K., "Simultaneous Feature Selection and Clustering Using Mixture Models", IEEE Transactions on Pattern Analysis and Machine Intelligence **26**, No. 9, 2004, pp. 1154-1166.

[12] Minaei-Bidgoli B., Topchy A., and Punch W., "Ensembles of Partitions via Data Resampling", in Proc. IEEE Intl. Conf. on Information Technology: Coding and Computing, ITCC04, **2**, April 2004, pp. 188-192.

[13] Topchy A., Jain A.K, Punch W., "A Mixture Model for Clustering Ensembles", Proceedings of the Fourth SIAM International Conference on Data Mining, Florida, USA, April 22-24, 2004.

[14] Topchy A., Jain A.K., Punch W., "Combining Multiple Weak Clusterings", In the third IEEE International Conference on Data mining, 19-22 June 2003, pp. 331-338.

[15] Topchy A., Minaei-Bidgoli B., Jain A.K., Punch W., "Adaptive Clustering Ensembles", Proceedings of International Conference on Pattern Recognition, ICPR'04, Cambridge, UK, 2004, pp. 272-275.

[16] Ujjwal Maulik and Sanghmitra Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices", IEEE Transactions on Pattern Analysis and Machine Intelligence, **24**, No. 12, December 2002, pp. 1650-1654.