ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 12, December 2016

Kernel Density based Germplasm Evaluation Using Cluster Ensemble

[1] Dr. Sangeeta Ahuja [2] Dr. H.L. Raiger [3] Dr. D. K. Ghosh
[1][4] Scientist (S.S.) Indian Agricultural Statistics Research Institute, New Delhi, India.
[2] Principal Scientist, National Bureau Plant Genetic Resources, India.
[3]Head and Professor, UGC Fellow, Saurashtra University, Gujarat, India.

***Abstract*:-- Germplam Evaluation plays very important role in genetic resources management and hybrid selection. The germplasm evaluation using the Kernel Density based Bayesian Ensemble (KDBCE) method gives promising results. It is based on Multivariate Kernel Density Estimation and Naive Bayes Classifier to obtain robust clustering using Density based (DBSCAN) clusterer. This algorithm operates in three phases. During the first phase, H input clustering schemes are generated by using the density based algorithm (DBSCAN) with different number of clusters in each clustering scheme. The optimum number of clusters is determined by computing the Silhouette coefficient for each clustering scheme. The second phase equalizes the number of clusters generated by different clustering schemes depending upon the optimum number of clusters. Accordingly, the clusters split or merge in different clustering schemes by using the kernel density based split and merge method. In the third phase, consensus partition is generated by the Naive Bayes Classifier. Empirical evaluation of the algorithm shows that the proposed method significantly improves the quality of resultant clustering scheme compared to the best of the original schemes.**

***Keywords:* DBSCAN, Cluster Ensemble, Naive Bayes Classifier, Cohesion, Separation, Silhouette Coefficient.**

I. INTRODUCTION

Getting the high quality results is very difficult by utilizing the traditional clustering methodologies because of the various anomalies viz., randomization, idiosyncrasies of algorithms [1], statistical, sampling techniques [2]. Germplasm evaluation have been done by traditional clustering methodologies and difficult to describe the best method as any one method. Cluster ensemble techniques [1, 3, 4, 5, 2, 6, and 7] can be utilized for combining multiple schemes for robust clustering solution. These methods not only improves the quality but also the accuracy and stability.

Problem Definition

Let D is a data set of N, d-dimensional vectors x =< x1,x2,...,xd >, each representing an object. D is subjected to a clustering algorithm which delivers a clustering scheme π consisting of K clusters. (π = {C1,C2,..., CK}). Let {π1,π2,...,πH} be H schemes of D obtained by applying either same clustering algorithm on D or by applying H different clustering algorithms. Further, let $\lambda\pi : D \rightarrow \{1,K\}$ be a function that yields labeling for each of the N objects in D. Let $\{\lambda_1,\lambda_2,...,\lambda_H\}$ be the set of corresponding labelings of D. The problem of cluster ensemble is to derive a consensus function Γ, which combines the H partitions (via labelings) and delivers a clustering π_f, with a promise that π_f is more robust than any of constituent H partitions and best captures the natural structures in D.

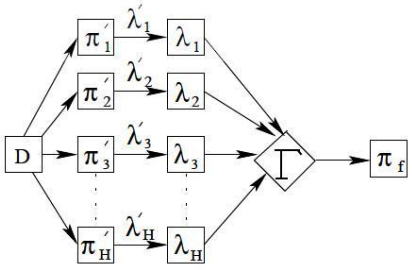

Figure 1: The process of Cluster Ensemble
Figure 1 shows the process of construction of cluster ensemble.

II DESIGN AND DEVELOPMENT OF KDBCE

KDBCE (Kernel Density Based Cluster Ensemble) has been developed in Windows environment with multithreaded Microsoft Visual C++ 6. 0 programming language which is an object-oriented language and suitable for software development. It utilizes various key features of object oriented technologies such as reusability, inheritability, encapsulation, portability and modular development. For details, a reference may be made to Kruglinski [9,14] and Ritcher. Discriminant Analysis and Logistic Regression have been performed

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 12, December 2016

using R 2. 0. 0 software distributed by CRAN Foundation (R Development Core Team, 2004).The Multilayer Perceptron has been performed using the data mining package WEKA Software (2000). The hardware used was a dual core Intel (R) machine (2. 20 GHz, 4 GB RAM). But any system with minimum of 64 MB RAM and 2 GB Hard Disk capacity is required to run this algorithm and it will be compatible on Microsoft Windows 98, ME, 2000 ,XP and . Net Platform.

III. ARCHITECTURE OF KDBCE ALGORITHM

The density based algorithm (DBSCAN) [13] for generating the original clustering schemes has been utilized in order to handle clusters of arbitrary shapes and sizes. KDBCE consists of three phases (see Figure 2). During the first phase, H input clustering schemes are generated by using the density based algorithm (DBSCAN) with different number of clusters in each clustering scheme. The optimum number of clusters is determined by computing the Silhouette coefficient for each clustering scheme. The second phase equalizes the number of clusters generated by different clustering scheme depending upon the optimum number of clusters. Accordingly, the clusters are splitted or merged in different clustering schemes by using the multivariate kernel density based split and merge method. In the third phase, consensus partition is generated by utilizing the Naive Bayes Estimation with Probability Index.
The KDBCE is mathematically shown as follows:

Let D denote the data set of N, d-dimensional vectors X_i =< X_1, X_2, . . . , X_d >, each representing an object. D is subjected to Density based clustering algorithm i.e. DBSCAN (Tan, et al. 2006). Let $\pi 1'$, $\pi 2'$, . . . , $\pi H'$ be H partitions of D obtained by repeatedly applying DBSCAN clustering algorithm with different parameters (ϱ and ρ). Each application results into different number of clusters in each partition, i.e. πi = (Ci1, Ci2, . . . , Ciki) where ki = number of clusters in the i^{th} partition.
Since, the number of clusters in each scheme may be different mapping of clusters in the scheme is not straight forward. It is necessary to first equalize the number of clusters in all schemes before proceeding. In the present context, this requires determining the optimal number of clusters before equalizing the number of clusters in H schemes. The optimum number of clusters is determined by computing the Silhouette coefficient for each clustering scheme and identifies the best clustering scheme. The number of clusters in this scheme is taken to be optimum, and to equalize the remaining schemes, clusters are merged or split, as the case may be. After equalizing the number of clusters by split/merge method the final consensus partition has been obtained.

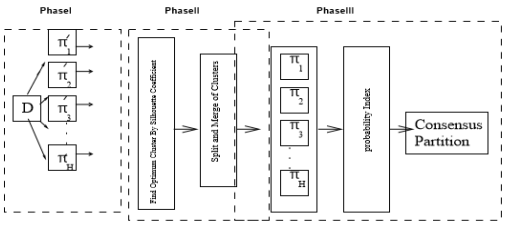

Figure 2: The architecture of KDBCE Cluster Ensemble

Experimentation
The experimentation has been executed on chickpea data set by implementing the KDBCE algorithm. The data has been collected from NBPGR, New Delhi, India. It consists of 3584 observations with 8 different characters viz., days of 50% flowering, number of primary branches per plant, plant height, number of pods per plant, number of seed per pod, days to 80% maturity, 100 seed weight, grain yield per plant and yield. Rigorous experimentation has been done by varying the cluster size and partitions for both the traditional and KDBCE cluster ensemble algorithm. Quality of KDBCE clustering and best of traditional clustering algorithm has been compared as shown in figure 3 and figure 4 respectively. In figure 3, 1 represents the purity of best of traditional clustering algorithm and 2 represents the purity by KDBCE algorithm. Similarly in figure 4, 1 represents the NMI of best of traditional clustering algorithm and 2 represents the NMI by KDBCE algorithm and results strongly improve the quality (Purity and NMI) by KDBCE cluster ensemble algorithm.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 12, December 2016

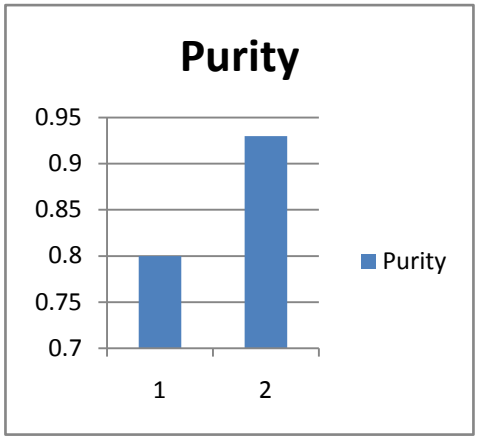

Figure 3: Purity comparison

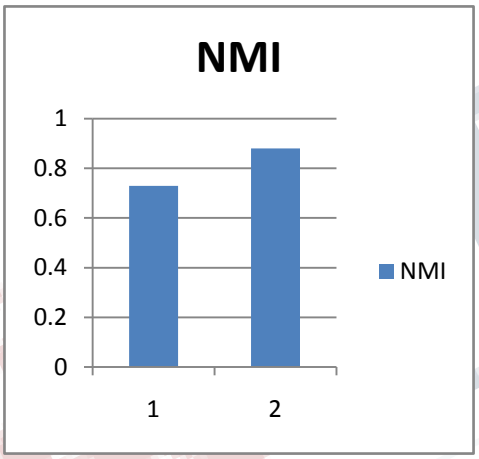

Figure 4: NMI comparison

Figure 5: Silhouette Coefficient

Other comparisons have been done for best of original (traditional) clustering scheme (Best OC) with the clustering scheme obtained from our cluster ensemble method that is, Kernel Density based Bayesian Cluster Ensemble (KDBCE) with optimum number of clusters and optimum partition. Best of Original clustering scheme is the partition at which we achieved the maximum silhouette coefficient value. 1 in Figure 5 refers to Best of silhouette coefficient and 2 refers to the KDBCE silhouette coefficient for chickpea germplasm. It clearly shows that through KDBCE cluster ensemble results improves a lot as compared to best of other clustering methods.

Assessing Quality of Cluster Ensemble

We employ the following measures for determining the quality of Performance Ensemble [14]

Purity (Accuracy):Let there be K clusters in the data set D corresponding to Kclasses in the data and the size of cluster C_j be $|C_j|$. If $|C_j|$class = idenotes number of objects of class i assigned toC_j , then purity of C_j isgiven as

$$\text{Purity}(C_j) = \text{Max}(i=1,K)\{|C_j|\text{class} = i/(|C_j|)\}$$

The overall purity of a clustering scheme is expressed as a weighted sumof individual cluster purity.

$$\text{Purity} = \Sigma_{(j=1,K)}(|C_j| * \text{Purity}(C_j))/|D|$$

In general, larger value of purity, indicates better quality of the solution.

- ***Normalized Mutual Information(NMI):*** Intuitively, the optimal combined clustering should share the most information with the original clusterings. Let A and B be the random variables described by the cluster labeling λ (a) and λ (b) with k (a) and k (b) groups respectively. Let I (A,B) denote the mutual information between A and B, H (A),H (B) denote the entropy of A and B respectively. Then normalized mutual information (NMI) is defined as follows

- NMI (A,B) = 2I (A,B)/H (A) + H (B)

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 12, December 2016

- Clearly, the value lies between [0, 1] and NMI (A,A) = 1.

- **Silhouette Coefficient (SC)**

Silhouette coefficient combines both Cohesion and Separation [8,11,12]. Cohesion measures the closely related objects in a cluster [8,11,12]. The steps for computing the silhouette coefficient for an individual point are as follows.

For the ith object, calculate the average distance to all other objects in its cluster. Call this value ai.

(i) For the ith object and a cluster not containing the object, calculate the object's average distance to all the objects in the given cluster. Find the minimum such value with respect to all clusters; call this value bi.

(ii) For ith object, the silhouette coefficient is given by

si = (bi-ai)/Max(ai,bi)

Silhouette coefficient for the clustering scheme is aggregation of the coefficients of individual points.

$$SC = \frac{1}{N}\sum_{i=1}^{N} s_i \qquad (3.4)$$

SC ranges [−1, +1], If SC = -1 represents bad, SC = 0 represents indifferent and SC = 1 represents good value.

IV. CONCLUSION

Kernel density based Germplasm evaluation by using the KDBCE Cluster Ensemble improves the results significantly as compared to other traditional clustering method. The algorithm of cluster ensemble consists of three phases from generation of clustering schemes to fusion of schemes in order to achieve the best final clustering scheme with compact clusters. The DBSCAN algorithm has been utilized for generation of initial clustering schemes. The unequal and irregular number of clusters can be handled by this KDBCE cluster ensemble. The advanced statistical techniques of Kernal Density Estimation and Bayesian concept have been used for construction of consensus partition. Empirical evaluation and extensive experimentation with chickpea data sets for germplasm evaluation shows better cluster quality with KDBCE as compared to best of original clustering schemes.

REFERENCES

[1] Reza Ghaemi, M., Nasir Sulaiman, H.I., Mustapha, N.: A survey: Clustering ensembles techniques. In: Proceedings of World academy of science, Engineering and Technology 38, 2070–3740 (2070).

[2] Topchy, A., Behrouz Minaei-Bidgoli, A., Punch, W.F.: Adaptive clustering en-sembles. In: ICPR, pp. 272–275 (2004).

[3] Kuncheva, L., et al.: Evaluation of stability of k-means cluster ensembles with respect to random initialization. IEEE Transactions on pattern analysis and ma-chine intelligence 11(28), 1798–1808 (2006)

[4] Fred, A.L.N., Jain, A.K.: Data clustering using evidence accumulation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 835–850 (2002)

[5] Topchy, A., Jain, A.K., Punch, W.F.: A mixture model for clustering ensembles. In: SDM (2004)

[6] Strehl, A., Ghosh, J.: Relationship-based clustering and cluster ensembles for high-dim. data. PhD thesis (May 2002)

[7] Fischer, B., Buhmann, J.M.: Path-based clustering for grouping of smooth curves and texture segmentation. Transaction on Pattern Analysis and Machine Intelligence 25(4) (April 2003).

[8] Jiawei Han and Micheline Kamber.- Data Mining : Concepts and Techniques Second Edition, Morgan Kaufmann Publishers, San Diego, USA, 2006.

[9] Kruglinski, David - Inside Visual C++, I Edition. Microsoft Press, Washington, 1996.

[11] Margaret H. Dunham - "Data Mining: Introductory and Advanced Topics", Southern Methodist University, Pearson Education Inc., Upper Saddle River, New Jearsey, 2003.

[12] Oded Maimon and Leor Rokech -. "Data Mining and Knowledge discovery Handbook", Springer publications, 2004.

[13] Richard A. Johnson and Dean W. Wichern, "Applied Multivariate Statistical Analysis", Prentice Hall, Upper Saddle River, New Jersey, 1979.

[14] Jeff , Prosise. Programming Windows with MFC, II Edition. Microsoft Press, Washington, 1999.