A COMPARATIVE STUDY OF CLUSTERING ALGORITHMS FOR WHEAT DATA

Dileep Kumar Yadav [#1], Maitreyee Dutta [*2]

[#]CSE Dept., VBS PU, Jaunpur, India

[*] Professor & Head, ECE Dept., NITTTR, Chandigarh

***Abstract*-Data mining is a promising research field in wheat data analysis. There are many data mining techniques such as clustering, classification, prediction, and outlier analysis can be used for the purpose of analysis. Clustering is a data mining technique used for discovering groups and identifying interesting distribution in the underlying data. Clustering algorithm such as k-means algorithm, density based, k-medoids and hierarchical based algorithms are used for various application in the data mining. There are several optimization methods are proposed in the literature in order to solve clustering limitations, but Swarm Intelligence has achieved its remarkable position in the concerned area. Particle Swarm Optimization is the most popular SI technique which is used in this paper. In this paper a comparative study is performed among k-means algorithm, Hierarchical clustering with centroid Linking and Correlation based feature selection with particle swarm optimization for clustering of wheat data set, to find out the best clustering algorithm in terms of accuracy rate in clustering the data.**

***Keywords*-Hierarchical clustering with centroid Linking, CFS with PSO, K-means, density based clustering**

I. INTRODUCTION

Data mining refers to as a field which deals with the search and research on the data. Mining is a term which means fetching or extraction of data from a large data set or we called as a huge data repositories. Data mining basically categorized into two types, Classification and Clustering. Both terms are different in nature from each other. Clustering firstly groups of objects can be prepared and then find out whether they relate with each other or not. Comparison of clustering for agricultural data is most challenging in this era. Agriculture is demographically the broadest economic sector and plays a significant role in the overall socio-economic structure of India. Agriculture as a business is unique crop production is dependent on many climatic, environmental, biological, political and economic factors that are mostly independent of one another.

II. DESCRIPTION OF DATASET:

The Dataset comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected from UCI machine learning repository for the experiment purposes [1]. High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology (M. Charytanowicz, *et. al*., 2010). The images were recorded on 13x18 cm X-ray KODAK plates.

The data set used in this paper contains seven geometric parameters of wheat kernels. These are described as follows:

1) Area A
2) Perimeter P
3) Compactness C = 4*pi*A/P^2
4) Length of kernel
5) Width of kernel
6) Asymmetry coefficient
7) Length of kernel groove

All of these are real-valued continuous parameters.

A. Normalization and standardization of data:

In the overall knowledge discovery process, before data mining itself, data preprocessing plays a crucial role. One of the first steps concerns the normalization of the data. This step is very important when dealing with parameters of different units and scales. For example, some data mining techniques use the Euclidean distance. Therefore, all parameters should have the same scale for a fair comparison between them. [10]

Two methods are usually well known for rescaling data. Normalization, which scales all numeric variables in the range [0, 1]. One possible formula is given below:

$$new = \frac{- \quad min}{max - \quad min}$$

On the other hand, you can use standardization on your data set. It will then transform it to have zero mean and unit variance, for example using the equation below:

$$new = \frac{- \mu}{\sigma}$$

B. Correlation Based Feature Selection:

CFS measures correlation between nominal features, so numeric features are first discredited.[7] However, the general concept of correlation-based feature selection does not depend on any particular data transformation—all that must be supplied is a means of measuring the correlation between any two variables. So, in principle, the technique may be applied to a variety of supervised classification problems, including those in which the class is numeric. CFS is a fully automatic algorithm—it does not require the user to specify any thresholds or the number of features to be selected, although

both are simple to incorporate if desired.CFS operates on the original (albeit discretized) feature space, meaning that any knowledge induced by a learning algorithm, using features selected by CFS, can be interpreted in terms of the original features, not in terms of a transformed space.

C. Particle Swarm Optimization:

The particle swarm optimization (PSO) algorithm is an optimization method developed by Eberhart.[2] PSO tries to find the optimal solution through the simulation of some ideas drawn from fish schooling, bird flocking, and other social groups. One such idea is that an agent can effectively achieve his objective using the information that is owned by him and the information that is shared among the group. This means that PSO is an optimization method that uses the principles of social behavior. PSO has proved to be competitive with genetic algorithms in several tasks, mainly in optimization areas. [11]

III. CLUSTERING OF WHEAT DATA BY USING HIERARCHICAL CLUSTERING WITH CENTROID LINKING:

K-means: K-means is a process of partitioning n-dimensional data into *k* sets to minimize the mean distance within each set. The most commonly used distance measures are the squared Euclidean distance and the sum of the squared differences across variables. Takes the input parameter, k, and partitions set of n objects into k cluster so that the resulting intra cluster similarity is high but the inter cluster similarity is low.

Hierarchical clustering: In data mining, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types: [4]

- Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Density based Clustering: It grows clusters according to the density of neighborhood objects. It is based on the concept of "density reachibility" and "density connect ability.[5]

X-means Clustering: The X-means algorithm starts with K equal to the lower bound of the given range and continues to add centroids where they needed until the upper bound reached.[6]

Hierarchical clustering with centroid linking: The description of the proposed method is described in Figure-1.

IV. RESULTS AND COMPARATIVE ANALYSIS:

Here, in this hierarchical clustering with centroid Linking is proposed to use with optimal features selected by correlation based feature selection with particle swarm optimization (CFS-PSO). Here, the result obtained by the proposed method is also compared with the k-means, Density Based Clustering with simple K-means, and x-mean with Manhattan distance. From the Obtained results it is observed that the proposed method perform better in comparison to other clustering techniques available in literature.

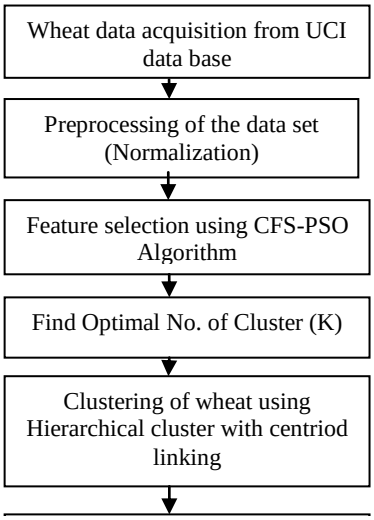

Figure 1: Methodology Flow Chart

V. PERFORMANCE MEASURE:

In this paper there is some measurement:

Accuracy: Accuracy is often the starting point for analyzing the quality of a predictive model, as well as an obvious criterion for prediction. Accuracy measures the ratio of correct predictions to the total number of cases evaluated. It may seem obvious that the ratio of correct predictions to cases should be a key metric. [8]

$$Accuracy = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{FP} + \mathrm{FN} + \mathrm{TP}}$$

Where,

TN is the number of true negative cases
FP is the number of false positive cases
FN is the number of false negative cases
TP is the number of true positive cases

Sum of squared errors: It is a measure of the discrepancy between the data and an estimation model. It is used as an optimality criterion in parameter selection and model selection. [9]

$$\mathrm{SSE} = \sum_{i=1}^{n} (\mathrm{yi} - \mathrm{f(xi)})^2$$

where *yi* is the i^{th} value of the variable to be predicted, xi is the i^{th} value of the explanatory variable, and f(xi) is the predicted value of *yi*

TABLE I
RESULT ANALYSIS FOR THE CLUSTERING OF WHEAT USING DIFFERENT CLUSTERING ALGORITHMS:

No. of clusters	DBCL-k-Means		k-Means		HCL with CL		X-Means with Manhattan	
	Full data	CFS-PSO data	Full data	CFS-PSO data	Full data	CFS-PSO data	Full data	CFS-PSO data
k	**Accuracy**	**Accuracy**	**Accuracy**	**Accuracy**	**Accuracy**	**Accuracy**	**Accuracy**	**Accuracy**
2	64.72	64.72	64.72	64.72	64.72	64.72	77.58	77.63
3	**88.05**	**86.15**	**88.05**	**88.20**	**90.95**	**87.58**	**86.15**	**88.10**
4	80.01	80.01	79.53	78.53	90.48	76.05	71.4	73.4
5	68.05	68.05	68.05	68.58	90.01	77.58	58.05	60.96

TABLE II
RESULT ANALYSIS FOR THE CLUSTERING OF WHEAT WITHOUT FEATURE SELECTIONS USING DIFFERENT CLUSTERING ALGORITHMS:

	k=2	**k=3**	k=4	k=5
DBCL-k-Means	64.72	**88.05**	80.01	68.05
k-Means	64.72	**88.05**	79.53	68.58
HCL with CL	64.72	**90.95**	90.48	90.01
X-Means with Manhattan	77.63	**86.15**	71.4	60.96

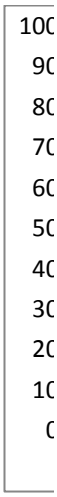

Figure 2: Graphical representation for the clustering of wheat without feature selections using different Clustering algorithms

TABLE III
RESULT ANALYSIS FOR THE CLUSTERING OF WHEAT WITH CFS-PSO FEATURE SELECTIONS USING DIFFERENT CLUSTERING ALGORITHMS

	k=2	**k=3**	k=4	k=5
DBCL-k-Means	64.72	**86.15**	80.10	68.05
k-Means	64.72	**88.05**	78.53	68.05
HCL with CL	64.72	**87.58**	90.48	90.01
X-Means with Manhattan	77.63	**88.10**	71.4	59.05

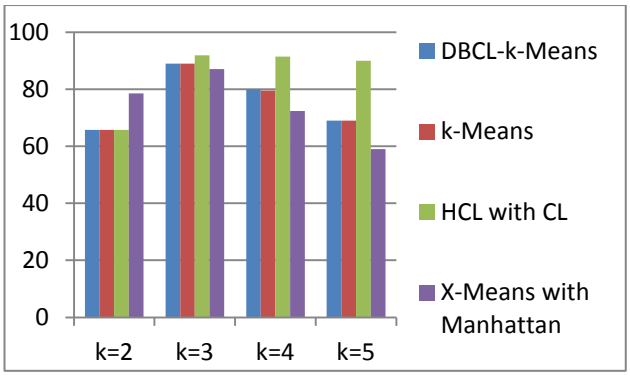

Figure 3: Graphical Representation for the clustering of wheat with CFS-PSO feature selections using different Clustering algorithms.

VI. CONCLUSION

Clustering of wheat is a most challenging problem in the agricultural field. In this paper various clustering algorithms have been used to cluster the wheat data such as density based clustering, k-means, hierarchal clustering with centroid linking and x-means with Manhattan distance. Here, correlation based feature selection with particle swarm optimization was used to select the optimal number of features. Here, it has been observed that the 86.15%, 88.05%, 87.58% and 88.10 % accuracy is obtained by density based clustering, k-means, hierarchal clustering with centroid linking and x-means.with Manhattan distance respectively. Here, from the result and analysis it has been observed that the hierarchal clustering with centroid linking provide the better results in comparison to the other clustering techniques available in literature.

VII. FUTURE WORK

A better clustering algorithm may be used to find accuracy and sum of squared error. This appears to happen when the data sets are very large. In future the performance of proposed method may be improved by extracting the other features using feature extraction techniques.

REFERENCES:

[1] Piotr Kulczycki, Malgorzata Charytanowicz, **"**A Complete Gradient Clustering Algorithm Formed With Kernel Estimators", Int. J. Appl. Math. Computer. Science., 2010, Vol. 20, No. 1, 123–134.

[2] Bighnaraj Naik, Subhra Swetanisha, Dayal Kumar Behera, Sarita Mahapatra, Bharat Kumar Padhi, "Clustering Algorithm based on PSO and k-means to find optimal clustec centroids," IEEE National Conference on Computing and Communication Systems 2012.

[3] Rehab F. Abdel-Kader, "Improved PSO Algorithm for Efficient Data Clustering," Second IEEE International Conference on Machine Learning and Computing, 2010.

[4] Nidhi Singh,.Divakar Singh,"Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time," International Journal of Computer Science and Information Technologies, Vol. 3 (3) , 2012,4119-4121.

[5] Aastha Joshi, Rajneet Kaur, " Comparative Study of Various Clustering Techniques in Data Mining," , International Journal of Advanced Research in Computer Science and Software Engineering 3(3), March - 2013, pp. 55-57.

[6] Dan Pelleg, Andrew Moore, "X-means Extending K means with efficient estimation of the number of clusters,"School of Computer Science, Carnegie Mellon University, Pittsburgh.

[7] Mark A. Hall, "Correlation-based Feature Selection for Machine Learning," Department of Computer Science the university of Waikato Hamilton, NewZealand.

[8] http://en.wikipedia.org/wiki/Accuracy_paradox.

[9] http://en.wikipedia.org/wiki/Residual_sum_of_square.

[10] http://www.dataminingblog.com/standardization-vs-normalization

[11] Ahmed A.A. Esmin1,2 and Stan Matwin,"Data Clustering Using Hybrid Particle Swarm Optimization," International Conference on Intelligent Data Engineering and Automated learning LNCS 7435, pp. 159-166, 2012.