



A Study of Clustering Techniques for Crop Prediction - A Survey

¹Utkarsha P. Narkhede, ²K. P. Adhiya

¹Research Student, ²Associate Professor

Department of Computer Science and Engineering

SSBT'S College of Engineering & Technology Bambhori, Jalgaon, Maharashtra, India

Abstract: Farming community necessitate for well organized system to predict and improve the crop over the world. The complexity of predicting the best crops is highly due to unavailability of proper knowledge discovery in crop knowledgebase which affects the quality of prediction. However, Clustering is an important step in mining useful information. There are several clustering methods such as partitioning, hierarchical, model-based, grid-based, constrained-based which make this task complicated due to problems related to optimization and noise. In this review paper there is a comparative study of clustering algorithms. Out of these BeeHive and Improved k-means clustering algorithm are outstanding in solving the optimization problem which led to select for performance evaluation in order to get good quality of clusters for crop prediction.

Keywords: BeeHive; Improved k-means; performance evaluation.

I. Introduction

The extent of details kept in computer files and databases is rising at an unexpected rate. At the equal period, the users of this text are look forward to get more refined information from them. A user is no longer delight with reference to simple listing of contents but wants information in dept about past as well predictions of future contents. On the level structure query languages are not satisfactory to support these increased demands of information. Hence steps of data mining are in used to solve these wants.

A. Concept of Data Mining

Data mining is a ground-breaking technology, developing with database and artificial intelligence. It is a processing overture of action of extracting trustworthy, novel, useful and understandable patterns from database. At current, Data mining has been in business management, production control, electronic commerce, market analysis and scientific research and many other fields to explore a wide range of applications. Data mining with a view of its socking business vision are now becoming an data library and information strategy-making in the field of agriculture research [1]. Data mining consist of functionalities that are used to specify the kind of patterns to be found in data mining task. There classification of data mining tasks is done in two categories, descriptive and predictive. Descriptive mining task characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions. In data mining, clustering plays an important role for finding data information and pattern recognition.

B. Clustering Analysis in Data Mining

Clustering is comes under the descriptive model that identifies patterns or relationship in data. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in agreement to one another but are very dissimilar to objects in option clusters. A cluster of data objects can be treated collectively during the time that one group and so may be considered as a classic of data compression. Unlike alignment, clustering is an effective means for partitioning the set of data into groups based on data similarity and then ascribe labels to the relatively small number of groups.

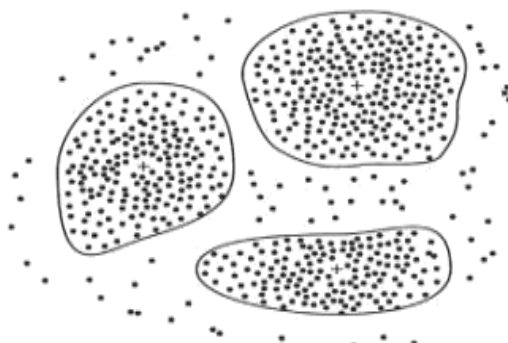


Figure 1: Cluster Analysis

During the time that a data mining appositeness, cluster dismemberment can be used as a stand-alone tool to possession insight into the distribution of data, to observe the characteristic of each cluster, and to focus on a particular set of clusters for further analysis. Clustering is an unsupervised learning as it does not rely on predefined classes and class-labelled training examples. For this reason, Clustering is a form of learning by observation, rather than learning by examples [2]. As shown in Figure 1, the three clusters are formed containing data points based on center position. The center are shown by + sign. The quality of clusters depends on how dense it is. So, cluster having more number of point is cluster of good quality.

C. Motivation

IT has added up to our daily living in earlier periods decade. Having IT enhancements in good organization can be made in about any branch of industry and packages, now days this become principally factual for agronomics. Now days farmer yield not only crops but along with rising amounts of records. These records are specific & small in range. On the other hand, assembling large amounts of records frequently is both a permission and irritation. There is a lot of data accessible containing information having absolute agreement. At this point such as properties of soil and yield should be used to the farmer's aid. Due to this data mining is needed. Data mining techniques such as clustering that aspire at finding those patterns or information in the data that are both worth full and motivating to the farmer.

A crop prediction is a widespread particular problem that occurs. During the rising season, a farmer had curiosity in knowing how much yield he is on the point of anticipate. In the earlier period, this yield prediction become a matter of fact relied on farmer's long-term experience for specific yield, crops and climatic conditions. Along with, the existing system never foresee points in focus as regards farmer economic status and their interaction, they directly away for yield prophecy rather than concerning crop prophecy because unless the good crop is predict how recur the yield will better, additionally pesticides, environmental & meterological parameter related to crop is not considered. Promoting and soothing the agricultural production at a more rapidly pace is one of the essential situation for agricultural improvement. Any crop's production show the way either by interest of domain or enhancement in yield or both. In India, the prospect of widening the district under any crop does not exist except by re-establishing to increased cropping strength or crop replacement. So, variations in crop productivity continue to trouble the area and generate rigorous distress. So, there is need to attempt good technique for crop prediction in order to overcome existing problem [3].

In this paper the clustering techniques are studied comparatively. The most efficient clustering techniques leading to accurate clustering of crop records are found out. The proposed system considers the input module, feature selection, crop database, two most efficient clustering approaches, pattern visualization and prediction rules and performance evaluation.

II. Literature Survey

To understand the advancement of clustering techniques, it is essential to briefly examine their history. In data mining, clustering plays an important role for finding data information and pattern recognition. Clustering is comes under the descriptive model that identifies patterns or relationship in data. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in agreement to one another but are very dissimilar to objects in option clusters. Hierarchical micro clustering algorithm, constrained k-Means algorithm, SWK k-Means algorithm, expectation maximization algorithm, improved k-Means algorithm are clustering techniques [4],[5],[6],[7].

A. Hierarchical Micro-Clustering Algorithm

Hwanjo Yu et.al presented clustering based support vector machine (CB-SVM) designed for handling very large data sets in 2003. Basically SVM is data classification method whose training complexity highly depends on size of data. So it is not worked for large dataset. In order to work authors have designed CB-SVM for handling large dataset. CB-SVM is a hierarchical micro clustering algorithm that scan entire data set only once to provide an SVM with high quality of sample that carry statistical summaries of the data.

Hierarchical micro-clustering algorithm introduces the clustering feature and clustering feature tree. Clustering feature is triple which summarize the information about cluster of objects such as number of data points, linear sum of data points and square sum of data points. Clustering Feature tree is a height balanced tree that stores the clustering features for a hierarchical clustering. The algorithm starts by scanning the database to build CF tree then SVM boundary function is trained from the centroids of the root entries of CF tree .The low margin clusters is determined that are close to the boundary and thus needs to be declustered into the finer level. Hierarchical micro-clustering algorithm shows good quality of cluster for large dataset but it is expensive to update and store the cluster and also splitting and merging the data degrades performance [4].

B. Constrained k-Means Algorithm

Kiri Wagstaff et.al developed HARVIST(Heterogeneous Agricultural Research Via Interactive, Scalable Technology) graphical interface that allows user to interactively run auto-matic classification and clustering

algorithm in 2005. They have used constrained k-means clustering algorithm for pixel clustering which merge the concept of constraint-based and partitioning methods.

In k-means instance alongside constraint are useful way to express a priori knowledge almost which instance should or should not be grouped together based on must-link and cannot-link constraints. It shows good quality of clusters for huge datasets and additionally give better activity than hierarchical clustering, but it has drawbacks such as the time that local optima problem, cannot find arbitrary shape cluster, backtracking to gain initial value of cluster feelings, affected by noise [5].

C. SWK k-Means Algorithm

A Majid Awan et.al has developed a software system in 2007 for predicting Oil-Palm Yield from climate and plantation data. They used unsupervised partitioning of data for finding spatio-temporal patterns using kernel method. By using only k-means partitioning method it is burdensome to deal with abstract data so authors have incorporated kernel method. Kernel exercise implicitly defines a non-linear transformation that maps the data from their original space to a high dimensional space where data are expected to be more.

A weighted kernel k-means clustering algorithm is used which incorporated spatial constraints to spare spatial neighbourhood information in order to handle noise. This algorithm requires added information such as kernel matrix, no of clusters, weight for each point, stopping criterion, penalty term parameter. It shows good quality of clusters for huge datasets and along with give better performance than hierarchical clustering, but it has global optima problem [6].

D. Expectation Maximization (EM) Algorithm

Sun Kim et.al proposed model for theme-based clustering algorithm that capture probabilistic for text documents in 2011. Probabilistic clustering comes under model-based clustering methods in which data are generated by mixture of probability distributions. Given text, subject terms are extracted and used for clustering document in a probabilistic framework. An EM algorithm is used for learning the proposed model in order to ensure annals are assigned to correct themes. EM algorithm is an iterative refinement algorithm used to estimation the parameters of the probability distribution so at the time that to best firm the data. It starts with initial estimates. Then, it iteratively refine the parameters based on Expectation and Maximization step. It is good in handling with real world dataset but becomes sensitive to noise and along with highly abstract in nature [7].

E. Improved k-means Algorithm

DUAN weing-ying et.al proposed improved k means clustering algorithm with weighted based on density in 2012. They proposed a solution to search initial central points and combines it with a distance measure with weight. This improved algorithm decreases the level of interference brought by the isolated points to the k-means algorithm, and makes the clustering analysis more effective and objective. An Improved k-Means clustering algorithm requires add-on parameter such as density, threshold and number of cluster. It is efficient in local as well as global optima and additionally reduces impact of noise data [8].

F. BeeHive Algorithm

M. Gunasundari et.al suggested crop yield prediction model which is used to predict crop yield from historical crop data set in 2013. A relational cluster Bee Hive algorithm is proposed for extracting yield patterns across multiple data sets. The outcome helps in identification of and investigates areas of unusually high or low yield. The BeeHive Algorithm is an optimization algorithm inspired by the natural foraging behaviour of honey bees to find the optimal solution. It searches for appropriate cluster centers such that the clustering metric is minimized. It requires the specific parameters to set such as scout bees, number of sites selected from neighborhood search, qualified sites, best site, number of bees recruited for the remaining selected sites and size of patch. The BeeHive algorithm is good in handling large dataset and efficient in local as well as global optima [1].

III. Literature Review

The comparison chart of different clustering algorithms is shown in Table 1. The chart is organized in terms of six algorithms such as hierarchical micro clustering algorithm, constrained k-Means algorithm, SWK k-Means algorithm, expectation maximization algorithm, improved k-Means algorithm and beehive clustering algorithm. And these algorithms are compared with the parameters like input parameter, optimization, dataset, shape and noise. In that, hierarchical micro clustering and constrained k-Means algorithm gives unsatisfied results by degrading performance in terms of optimization, EM and SWK k-Means algorithm are efficient in only local optima but not in global optima but improved k-Means and BeeHive clustering algorithms are good enough to deal with local as well as global optima problem. Excluding BeeHive algorithms, all algorithms formed their cluster in spherical shape whereas beehive formed hexagonal shaped cluster. Except EM algorithm, all algorithms are good enough to deal with noise.

Table 3.1 Comparison Chart of Different Clustering Algorithms

Algorithms	Input Parameter	Optimization	Dataset	Shape	Noise
Hierarchical Micro Clustering	Branching factor, Diameter threshold	Performance degrades	Large Dataset	Spherical	Yes
Constrained k-means	Must-Link, Cannot-Link, Number of clusters	Inefficient in local as well as global optima	Small Dataset	Spherical	No
SWK k-means	Kernel Matrix, Number of clusters, weight, penalty term	Efficient in only local optima	Large Dataset	Spherical	Yes
EM	Number of clusters	Efficient in local optima but not in global optima	Real world Dataset	Spherical	No
Improved kmeans	Density, Threshold, Number of clusters	Efficient in local as well as global optima	Huge Dataset	Spherical	Yes
BeeHive	scout bees, number of sites selected from neighborhood search, qualified sites, best site, number of bees recruited for the remaining selected sites and size of patch.	Efficient in local as well as global optima	Large Dataset	Hexagonal	Yes

IV. Proposed System

In this paper the authors have studied clustering techniques such as Hierarchical micro clustering algorithm, constrained k-Means algorithm, SWK k-Means algorithm, expectation maximization algorithm, improved k-Means algorithm. But this make the task complicated due to several problems related to optimization & noise. The initial perception fails to take into account the factor such as Global optima, No proper decision of initial cluster center, Accuracy, Time. If there is continuation in believing on this theory then the larger question of optimization is never understand and it will affect the quality of cluster. With reference to literature work & problem statement, clustering is the area of concern. All clustering algorithm are implemented for different dataset. Out of these, two algorithms BeeHive & Improved kmeans are outstanding in solving optimization problem & also good in outlier handling. Both algorithms have overcome the problems of k means, EM, Hierarchical clustering algorithms. So these algorithms are chosen for performance evaluation in crop prediction in order to get good quality of cluster.

Architecture of Crop Prediction

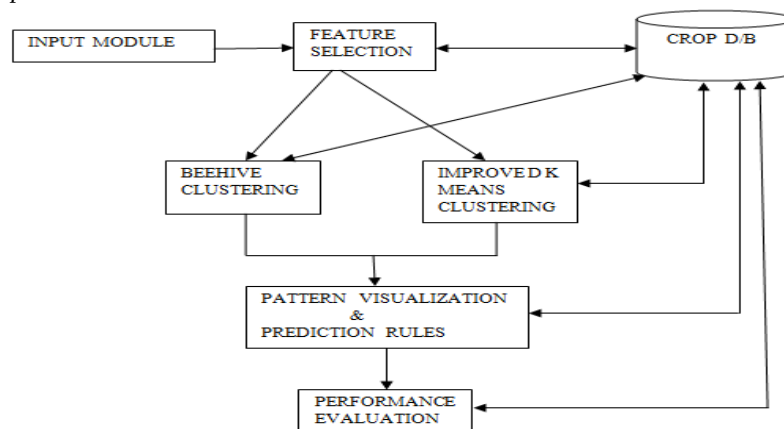


Figure 4.1 Architecture of Crop Prediction

Figure 4.1 shows the architecture of crop prediction which includes an input module which is responsible for taking input from farmer. In that the farmer has to provide area of land, region, economic status and city. The farmer is also responsible for interacting with predicted results. After selecting the city parameter based on altitude, longitude and latitude automatic climatic data will be reflected from crop knowledge base. The feature selection module is responsible for subset selection of attribute from crop knowledge base for robust learning. The crop knowledge base is consist of farm knowledge such as region-id, region-name, soil-type, water ph, rainfall, humidity, sunlight, land information, environmental parameter, city, pesticides information, crop knowledge such crop type, seed type. The knowledge-base also includes the samples of crop with corresponding farm knowledge, environmental parameter, and pesticides information. After subset selection of attribute, the data goes to two clustering approaches for grouping similar contents. The clustering approaches are BeeHive and Improved k-Means algorithm. These clustering approaches can form group of highly similar data point which is observed by pattern visualization. Then prediction rules will be applied to output of clustering to get

results in terms of crop, pesticide and cost. The performance is evaluated between the provided samples in crop knowledge base and predicted results.

V. Conclusion

As the mining of the most relevant information is important for the accurate crop prediction, authors have comparatively studied various clustering techniques in this paper. Out of this clustering algorithm, BeeHive and Improved k-means have solved the optimization problem and they are also good in outlier handling. The proposed method considered both algorithms which can be used to find out the most efficient and accurate crop from obtained clusters. So, the performance evaluation of both algorithms can be helped to know best clustering algorithm in terms of efficiency, accuracy and time.

References

- [1] M. Ananthara, T. Arunkumar, and R. Hemavathy, Cry: An improved crop yield pre-diction model using bee hive clustering approach for agricultural data sets," in *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2013 IEEE International Conference*, pp. 473-478.
- [2] J. Han, M. Kamber, and J. Pei, *Data Mining, Second Edition: Concepts and Techniques*, ser. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2006. [Online]. Available: <http://books.google.co.in/books?id=AfL0t-YzOrEC>.
- [3] R. A.A. and K. R.V., Review - role of data mining in agriculture," *International Journal of Computer Science and Information Technologies(IJCSIT)*, vol. 4(2), no. 0975-9646, pp. 270-272, 2013.
- [4] J. Y. Hwanjo Yu and J. Han, in *Classifying Large Data Sets Using SVMs with Hierarchical Clusters*. ACM New York, NY, USA 2003, pp. 306-315.
- [5] D. M. Kiri L. Wagstaff and S. R. Sain, "Harvist: A system for agricultural and weather studies using advanced statistical methods," 2005. [Online]. Available: <https://www.agriskmanagementforum.org>
- [6] A.Awan and M.Md.Sap, "A framework for predicting oil-palm yield from climate data", *International Journal of Information and Mathematical Sciences*, 2012.
- [7] S. Kim and . Wilbur, "An em clustering algorithm which produces a dual representation," in *Machine Learning and Applications and Workshops (ICMLA), 2011 IEEE 10th International Conference on*, vol. 2, pp. 90-95
- [8] W.-y.Duan, -r. Qiu, L.-z. Duan, Q. Liu, and H.-q. Huan, "An improved rough k-means algorithm with weighted distance measure," in *Granular Computing (GrC), 2012 IEEE International Conference*, pp. 97-101.

Acknowledgments

The authors feel a deep sense of gratitude to Prof. Dr. G. K. Patnaik HOD of Computer Science and Engineering Department for his motivation and support during this work. The authors are also thankful to the Principal, SSBT's College of Engineering and Technology Bambhori, Jalgaon for being a constant source of inspiration.