Application of Cluster Analysis in Agriculture – A Review Article

Mamta Tiwari
Dept. of Computer Application,
U.I.E.T.,C.S.J.M. University,
Kanpur, India

Dr. Bharat Misra
Dept. of Physical Sciences,
M.G.C.G. Vishwavidyalaya,
Chitrakoot. (M.P.) India.

ABSTRACT

In this paper we humbly present a review of some applications of cluster analysis in the field of agriculture and allied sciences. A few applications among them, which have been discussed here includes hierarchical agglomerative clustering approach, fuzzy clustering, hierarchical divisive clustering and Kohonen self-organizing feature maps along with an application of each of these techniques in the field of agriculture is also presented. Data mining in agriculture is a relatively new research field and the use of cluster analysis has almost been just begun in this area. This is our strong belief that effective techniques can be carved out for solving different agricultural problems of various complexities by intelligent use of data mining and its tools such as cluster analysis.

Keywords

Agriculture, Data mining, Fuzzy clustering, Hierarchical agglomerative clustering, Hierarchical divisive clustering, Kohonen self-organizing feature maps.

1. INTRODUCTION

This is an era of knowledge and information. There is virtually an explosion of information these days. Agriculture is considered to be the oldest profession of mankind. Due to several vagaries, related to climate, pests and others, the condition of agriculture was getting in shambles. The use of computers in the field of agriculture thus creates a very pleasant scenario where one of the most recent technologies comes forward in aide to one of the oldest inventions of human race. One very major task that has been evolved now a day is to mine an agricultural knowledge base.

Agricultural expert systems are being used extensively almost in every walk of life. Various tools also have been evolved for evaluating, justifying, upgrading and modifying the existing agricultural expert systems thus making them more useful in their intended purposes. The current paper reviews the use of cluster analysis as a tool of improving agricultural management, predicting and suggesting solutions for its problems and briefly discusses few such efforts.

Before discussing applications of data mining especially cluster analysis in the field of agriculture, let us have a look on what clustering is and various methods and techniques used for clustering.
Clustering is the process of grouping or making sets of similar or nearly similar type of physical or abstract objects. The groups thus formed are known as clusters. It is the process of grouping the data into classes or clusters, so that the objects within the same cluster have higher degree of similarity in comparison to one another but are very much dissimilar to the objects in different clusters [1]. We can compare the clusters with classes as in object-oriented programming paradigm. The slight difference between cluster and class is that, in class every object of it is exactly identical in properties whereas in cluster, every object is almost similar to other objects of its cluster and on the other hand, dissimilar to the objects of other cluster, if comparison is done on the basis of some particular properties of the objects.

There are several clustering techniques available and those are organized into the following categories as partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods, methods for high-dimensional data and constraint-based clustering. We here, however limited our discussion to hierarchical agglomerative clustering, fuzzy clustering, hierarchical clustering and Kohonen self-organizing feature maps only because these are the widely used data mining methods in the field of agriculture and allied science.

This paper has been organized as following. In section 2, we discussed the above written clustering methods. Subsection 2.1 particularly deals with hierarchical agglomerative clustering. Subsections 2.2, 2.3 and 2.4 are associated with fuzzy clustering, hierarchical divisive clustering and Kohonen self-organizing feature maps respectively. In the subsections of section 3, we discussed various applications of these methods in the field of agriculture or related fields in the above written order. A brief summary and future scope of application of cluster analysis in the discussed field is given in section 4. Following this section, we conclude the paper with used references in section 5.

2. CLUSTERING METHODOLOGY

We here briefly present, in order, the above stated methods of clustering techniques.

2.1 Hierarchical Agglomerative Clustering

The classic example of hierarchical agglomerative clustering is species taxonomy. The hierarchical agglomerative approach which is also known as the bottom-up approach starts by placing each object in its own cluster. The next step is to merges these atomic clusters into successively larger clusters, until all of the objects are confined in a single cluster or until certain termination conditions are satisfied. Most hierarchical clustering methods belong to this category. They differ only in their definition of inter-cluster similarity [2].

2.2 Fuzzy Clustering

The above discussed partition clustering method mainly deals with the task of partitioning a set of entities into a number of homogeneous clusters, with respect to a suitable similarity measure. That is also known as hard clustering. In other words, in hard clustering, the data element is divided into distinct clusters, where each data element belongs to exactly one cluster and we can predict the association of any data element to the cluster just by knowing that data element's that particular property on the basis of which the partitioning has been done. Many practical problems may also have fuzzy nature and because of their fuzzy nature, a number of fuzzy clustering methods have also been developed, following the general fuzzy set theory strategies developed by Lotfi Zadeh [3].

In fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one clusters simultaneously, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster or in other words, the participation of that particular data element with any particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters [4]. The main difference between the traditional hard clustering and fuzzy clustering can be stated as follows. While in hard clustering an entity belongs only to one cluster, whereas in fuzzy clustering entities are allowed to belong to many clusters with different degrees of membership. Among several available algorithms, one of the most widely used fuzzy clustering algorithms is the Fuzzy C-Means (FCM) Algorithm (Jim Bezdek 1981).

The FCM algorithm attempts to partition a finite collection of n elements, $X = \{x_1,...,x_n\}$ into a collection of c fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of c cluster centers where $C = \{c_1,...,c_c\}$ and a partition matrix $U = u_{i,i} \in [0,1]$, $i = 1,\ldots,n$ and $j = 1,\ldots,c$ where each element u_{ii} tells the degree to which i element belongs to cluster c_i. Like the k-means algorithm, the FCM aims to minimize an objective function [5].

2.3 Hierarchical Divisive Clustering

A hierarchical method creates a hierarchical decomposition of the given set of data objects. It can be classified as being agglomerative, as discussed earlier or divisive, based on how the hierarchical decomposition is formed.

The divisive hierarchical clustering approach, which is also known as the top-down approach, starts with all of the objects within the same cluster. In successive iteration, a cluster is split up into several smaller clusters, until eventually each object is placed in its own cluster, or until a termination condition holds [2].

2.4 Kohonen Self-Organizing Feature Maps

Self-organizing feature maps (SOMs) are one of the most popularly used neural network methods for cluster analysis. Kohonen networks were introduced in 1982 by Finnish researcher Tuevo Kohonen. Although applied initially to image and sound analysis, Kohonen networks are an effective mechanism for clustering analysis. Kohonen networks represent a type of Self Organizing map (SOM), which itself represents a special class of neural network [2].

The goal of SOM is to convert a high dimensional input signal into a simpler low dimensional discrete signal. In SOM, a set of nodes is arranged in geometrical pattern. SOM is an algorithm that is inspired by neural network in brain and that forms clusters by mapping high dimensional data into a 2-D or 3-D feature map. SOMs' goal is to represent all points in a high-dimensional source space by points in a low-dimensional (usually 2-D or 3-D) target space, such that the distance and proximity relationships (and hence the topology) are preserved as much as possible [2].

With SOMs, clustering is performed by having several units competing for the current object. The unit whose weight vector is closest to the current object becomes the winning or active unit. So as to move even closer to the input object, the weights of the winning unit are adjusted, as well as those of its nearest neighbours. SOMs assume that there is some topology or ordering among the input objects and that the units will eventually take on this structure in space. The organization of units is said to form a feature map. SOMs are believed to resemble processing that can occur in the brain and are useful for visualizing high-dimensional data in 2-D or 3-D space [2].

3. APPLICATIONS IN THE FIELD OF AGRICULTURE

3.1 Application of Agglomerative Clustering

The use of hierarchical agglomerative clustering is nicely depicted by Georg Ruß, Martin Schneider and Rudolf Kruse for management zone delineation in precision agriculture.

Precision agriculture is mainly concerned with the use of technology and the integration of various technologies with agriculture. As the result of advancement in the field of science and technology, the cost of technology is getting down day by day. Furthermore this technology is getting embedded into various agricultural equipments also. As a result of this integration, agricultural equipments of present time become more productive, updated and useful for farmers.

This also results in flooding of information, generated by these equipments as GPS crop growth sensors, fertilizer usage sensors and high-resolution satellite or aerial imaging. These sensors generate spatial data sets. Therefore, methods that take these special properties into account have to be developed to cope with the tasks encountered in precision agriculture [6].

One of these tasks is the delineation of management zones. This process is usually required before the growing season, when the availability of different mineral as Potassium, Phosphorus and Magnesium has to be measured and must be made available, termed as base fertilization, as it may be vital for the healthy growth of the crop. The delineation of management zones has been used as a method of subdividing fields into parts with different properties for a long time. However, this has usually been done earlier using expert and long-term knowledge of the respective field. In their approach, Georg Ruß, Martin Schneider and Rudolf Kruse developed a two step process.

The first step of spatially partitioning the data points may be achieved by overlaying a grid. Due to the irregularities in the field shape and the gaps as well as holes in natural data density, running a k-means algorithm on the coordinates of the points in the data set provides a more flexible solution to the initial tessellation. An upper bound for the parameter k is given by the size of the resulting smallest zone while zones below a threshold that is being provided by the precision of the used farming equipment cannot be managed. A lower bound for the k parameter is set by the granularity of the final management zones and by the amount of heterogeneity on the field [6].

The second step of repetitively merging of two zones has two constraints: first, zones which are to be merged must be similar in their attributes; second, they must be direct neighbours in geographical space (spatial constraint). As the consequence of both the conditions, it would be ensured that the resulting zones will rather be homogeneous, according to the first condition and contiguous, according to the second condition.

The spatial constraint can easily be fulfilled by generating a list of neighbours for each cluster and updating this list accordingly on cluster merge steps. The similarity measure, that decides which of the spatially neighbouring clusters are to be merged is usually one of single-linkage, complete linkage or average-linkage. Single-linkage would determine the data records which are most similar in two candidate clusters and merge the clusters containing these records. Complete-linkage would determine those data records which are most dissimilar in candidate clusters and merge those clusters. Average-linkage would determine the average vector of the data records in one cluster and would then merge those clusters which are closest according to the (usually Euclidean or Cosine) attribute distance. It is assumed that an average vector characterizes a (spatial) cluster sufficiently well [6].

3.2 Application of Fuzzy Clustering

Mohammad El-Helly, Hoda Onsi, Ahmed Rafea and Salwa El-Gammal in their study used fuzzy clustering in detection of leaf spots in cucumber crop [7]. Leaves spots are indicative of plant diseases; earlier leaf batches are examined manually and are then subjected to expert advice. The experts after the proper investigation, used to announce the disease. In their effort, Mohammad El-Helly, Hoda Onsi, Ahmed Rafea and Salwa El-Gammal, proposed a segmentation technique for identifying leaf batches in cucumber crop, based on fuzzy clustering algorithm.

The first step of image analysis and pattern recognition is the segmentation of image. Segmentation can be viewed as a clustering problem. It is very critical and inevitable component of image analysis and pattern recognition. This is the task which determines the quality of image analysis. Image segmentation is carried out by partitioning the image into homogeneous disjoint regions pertaining to some criterion as intensity or colour and none of the union of any two adjoining region should be homogeneous. The segmentation techniques distinguish the region with certainty but it may not be the case always. The regions in an image may not be defined very precisely always and there may be some uncertainty at each level of image processing process.

There are some mechanism provided by fuzzy set theory to represent and manipulate the ambiguity and uncertainty. Fuzzy set theory provides a function that provides a natural means to model the ambiguity or uncertainty in any image. In traditional clustering techniques, here are only two binary values, either 0 or 1 to determine the belonging or association of a data point to a cluster. The cases in real world for image analysis particularly are quite different than this, where the boundaries between clusters are not always clearly defined but there may be overlapping of grey scale intensities. In particular in the case of plant images, borders between tissues are not well defined and membership in the boundary region in essentially fuzzy in nature. Thus fuzzy clustering turns out to be particularly suitable for segmentation of plant images. That is the reason, Ahmed Rafea et al., have used fuzzy c-mean algorithm for this process.

They took 90 images of defected cucumber leaves to cover mainly disorder of three categories, namely powdery mildew, downy mildew and leafminer. Then fuzzy c mean algorithm was applied. The success of the application of this algorithm depends highly upon adapting the input parameters. The parameters were including the feature of data set, the optimal number of cluster and the degree of fuzziness. The feature of data set that considered was, mainly the image intensity because the defected portion of the leaves had high intensity comparing to the other healthy parts of the leaves. The intensity image is defined as the average value of red, green and blue colours. Thus the clusters having high intensity center represents the defected part of a leaf. The x and y coordinates were taken as the two other features for spatial information.

As for the optimal number of cluster is concerned, three measure out of several had been chosen. These were partition coefficient, partition entropy and compactness and separation. Partition coefficient measures the closeness of the samples to their respective cluster center.

Whereas the partition entropy measures the average amount of information contained. As the membership value gets deviated from the two extremes ie. 0 and 1, which represents good clustering results and hence small entropy, the entropy gets greater and as the membership value gets closer to the mid way that is 0.5 the entropy attains its maximum value. This is the indication of poor clustering results and high degree of fuzziness.

The last used measure, that was compactness and separation, represented the ratio of the average distance of the input samples to their corresponding cluster center and the minimum distance between cluster centres. Good cluster result should bring all input samples to as close as possible to their respective cluster and the different cluster centres as far as possible.

The degree of fuzziness is decided by fuzzy exponent factor ‘m’. The performance of FCM algorithm depends on a good choice of it. There are different views of several researchers regarding the value of this parameter. The range taken in this study by Ahmed Rafea et al. was {1.1, 1.5, 2, 2.2, 3, 8, 15}. As a result of applying FCM algorithm on the selected data set, they came to the conclusion that the optimal number of cluster is 4 whereas the degree of fuzziness is 2 for leaf spot problem. This set of values gives almost accurate result for segmentation of leaf spots.

Thus FCM algorithm emerges as the most suitable algorithm for detection of diseases in plants that can be sensed by inspecting the leaves. Although this study was aimed for cucumber plant, nevertheless this method can be expanded for other crops also.

3.3 Application of Divisive Clustering

The climate of Iran is characterized by complex pattern of spatial and temporal variability, with wide unpredictable rainfall fluctuations that varies from year to year and from region to region. Therefore, it is difficult to know the regional variation of rainfall. Divisive Clustering is being used by Saeed Soltani and Reza Modarres in their study of rainfall pattern and its classification in Iran [8]. This study is not directly related to any crop in particular, rather to agriculture as a whole. When there cannot be any certain prediction of the rainfall, in that case, the identification of rainfall pattern becomes an essential task for regional and local planners and managers. It has always been a concern for hydrologists to classify hydrologic events in order to simplify hydrologic convolution and thus reduce time and save budget for their planning and strategies. Multivariate techniques have been underlined as suitable and powerful tool to find homogeneous region on the basis of rainfall or to classify meteorological data such as rainfall. Principle component analysis, factor analysis and different cluster techniques have been used to classify daily rainfall patterns and their relationship to atmospheric conditions (Romero et al. 1999).

The data set used by Saeed Soltani and Reza Modarres in this study includes annual and seasonal rainfall of 28 capitals of their provinces of Iran. The variation of the mean rainfall of Iran is 224 mm to 300 mm which is widely distributed over Iran. Firstly the descriptive statistics of the annual rainfall was calculated, that includes mean, standard deviation, coefficient of variation (Cv), coefficient of skewness (Cs) as the measure of symmetry and coefficient of kurtosis (Ck) as the measure of shape of frequency function.

The first attempt of analysis of the clusters was done using k-mean. It was found that the number of suitable cluster seems to be 3, 6 or 8. These values show the possibility of the existence of 3, 6 or 8 regions based on rainfall regimes but could not be determined the regions by *K*-means algorithm. Then hierarchical methods were applied to find out the suitable number of the clusters.

Based on Euclidean distance, two method of hierarchical clustering, average method and Ward method were used to classify annual rainfall into 12 similar clusters. Pseudo *F* and t^2 was calculated for these clusters which are another suitable method to select the number of clusters besides minimum error function. Both pseudo F and t^2 had significant changes at 3, 6 and 8 clusters which indicate potential number of clusters for classifications. The coefficient of determination (R^2) increases with the number of clusters. The values of R^2 of two, three, six and eight clusters were 0.632, 0.804, 0.930 and 0.96, respectively. In this case, they accepted 8 clusters which cover more than 95% of rainfall variance over Iran. The result of this analysis was usually shown in a graphical illustration called “Dendrogram”.

In order to illustrate these clusters, they applied Canonical Discriminant Analysis (SAS/STAT, 1999). Canonical discriminant analysis is a dimension-reduction technique related to principal component analysis and canonical correlation. In a canonical discriminant analysis, they found linear combinations of the quantitative variables that provide maximal separation between the classes or groups. This study was aimed to classify annual rainfall over Iran into spatial groups. It was found that a hierarchical cluster analysis could classify this spatial pattern. The comparison of derived clusters and geographical conditions was very well matched with each other.

3.4 Application of Kohonen Self-Organizing Maps

The application of the clustering methodology, namely Self-Organizing Maps (SOM) is nicely done in the field of rice productivity by Shafaatunnur Hasan and Mohd Noor Md Sap [9]. Rice is considered to be an integral part of the food for most of the population of the world. An increase in productivity of this grain is always admired.

In their effort, they tried to conquer the battle against the rice pests with the help of natural enemies of rice pests as parasite, predators and pathogens, taking help of cluster analysis tools, particularly Kohonen Self-Organizing Maps (SOM). The losses incurred to rice crop due to insects, birds and rats are estimated to be between whopping 10% - 15% [10]. They presented an intelligent solution by implementing spatial analysis and Kohonen Self Organizing Map to cluster, types of pests, for better agricultural rice pest management in Malaysia.

The SOM consists of a regular, usually two-dimensional grid of map units. Each unit *i* is represented by a prototype vector *mi = [mi1,….,mid]*, where *d* is the dimension of input vector. The number of map units, which may vary from a few dozen up to several thousand, determines the accuracy and generalization capability of the SOM. Data points lying near each other in the input space are mapped onto nearby map units. The SOM is trained iteratively. At each training step, a sample vector *x* is randomly chosen from the input dataset. Distances between x and all the prototype vectors are computed. The best matching unit is the map unit with prototype closest to *x*, that is ||x-*mi*||=min{||x-*mi*||}. Next, the prototype vectors are updated. The best matching unit and its topological neighbours are moved closer to the input vector in the input space.

A SOM was trained using the sequential training algorithm for each data set. All maps were linearly initialized in the subspace spanned by the two eigen vectors with greatest eigen values computed from the training data. The maps were trained in two phases: a rough training with large initial neighbourhood width and learning rate and fine-tuning phase with small initial neighbourhood width and learning rate. The neighbourhood width decreased linearly to 1 with Gaussian function.

The period of data in the study was taken from 1996 to 1998 with 4 areas and 27 locations. With two planting season for each year, a total of 6 seasons were generated. There were 35 parameters that affected the rice yield. They classified these parameters in 5 groups. These include: three types of weed which are *rumpai, rusiga and daun lebar*; Three types of pests: rats, type of worms and *bena perang*; Three types of diseases: bacteria *(blb & bls), jalur daun merah* and *hawar seludang;* one type of lodging and one type of wind paddy. From 35 parameters, only 11 parameters are chosen since these are the most significant features as suggested by the experts of the respective field.

In their study, SOM network with 2 Dimensional and 10x10 lattice square neuron was applied with 27 observations, 11 variables, 10 neurons, 1000 times learning cycle with learning parameter from 0.9 to 0.1 and Gaussian Neighbourhood as percentage map width that started from 50 and reducing to 1. In this experiment, the learning parameter and Gaussian Neighbourhood was used as Exponential Decay to shrink SOM’s lattice structure.

These results have proven that pests such as rats, type of worms and *bena perang* are one of the factors that have affected the rice production. Pests and weeds are the major factor of the rice yield losses in Malaysia. Hence, intelligent solutions are needed to mitigate the issues of rice productivity. As such, intelligent clustering that is based on SOM network has been successfully applied in spatial analysis for Integrated Pest Management (IPM) in Malaysia [9].

4. SUMMARY AND FUTURE SCOPE

In the countries of the third world where proper facilities for irrigation, proper distribution of fertilizers, proper management, conservation and storages etc. are not available, furthermore almost entire agriculture and in turn, the economy, primarily depends on the amount of rainfall, the successful application of the study such as done by Saeed Soltani and Reza Modarres can bring a mammoth change. We strongly believe that data mining and cluster analysis should be a part of agriculture because they can improve the accuracy of decision systems. The cluster heuristic allows data to be combined into useful patterns that may lead to better decisions.

In present scenario the application of cluster analysis has already gained momentum, still there are lot of areas where a great deal of efforts is still required. The knowledge engineers and information scientists have done tremendous work in form of knowledge bank and knowledge processing; now there is an immense need to upgrade that work and make that even more useful. We believe that various data mining approaches and techniques such as k-mean, pCluster and STING etc. are going to play a vital role in future in this mega job.

5. REFERENCES

[1] Han, J., Kamber, M.: Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann (September 2000)

[2] Jiawei Han, Micheline Kamber: Data Mining Concept and Techniques 2nd Ed. - Morgan Kaufmann Publishers.

[3] A Fuzzy Clustering Model of Data and Fuzzy c Means; S. Nascimento, B. Mirkin and F. Moura Pires

[4] "Cluster analysis" in http://public.fh-wolfenbuettel.de/~hoeppnef/clustering.html

[5] "Fuzzy Clustering" in http://en.wikipedia.org/wiki/Fuzzy_clustering

[6] Hierarchical Spatial Clustering for Management Zone Delineation in Precision Agriculture: Georg Ruß, Martin Schneider, Rudolf Kruse

[7] Detecting Leaf Spots in Cucumber Crop Using Fuzzy Clustering Algorithm: Mohammed El-Helly, Hoda Onsi, Ahmed Rafea, Salwa El-Gammal

[8] Classification of Spatio -Temporal Pattern of Rainfall in Iran Using A Hierarchical and Divisive Cluster Analysis: Saeed Soltani and Reza Modarres, Journal of Spatial Hydrology Vol.6, No.2 Fall 2006

[9] Pest Clustering With Self Organizing Map for Rice Productivity: Shafaatunnur Hasan and Mohd Noor Md Sap

[10] MARDI, "Manual Penanaman Padi Berhasil Tinggi Edisi 1/2001", Malaysian Agriculture research and Development Institute. Malaysian Ministry of Agriculture (2002).