www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 4 April, 2014 Page No. 5369-5372

Identification of diseases of Soybean Using Cluster Analysis

Mamta Tiwari, Dr. Bharat Mishra

Dept. of Computer Application,
U.I.E.T., C.S.J.M. University, Kanpur, India
themamta1@gmail.com
Dept. of Physical Sciences,M.G.C.G. Vishwavidyalaya,
Chitrakoot. (M.P.) India.
bm_cgv@rediffmail.com

Abstract

In this paper we humbly present an effort in the direction of identifying major diseases in soybean crop, by application of cluster analysis. Data mining in agriculture is in itself a relatively new research field. The use of cluster analysis in the field of agriculture is relatively in its nascent form. A few techniques of cluster analysis have also been discussed here. It includes hierarchical agglomerative clustering approach, fuzzy clustering, hierarchical divisive clustering and Kohonen self-organizing feature maps. This is our strong belief that more effective techniques can be carved out in future for finding the root cause factors for spreading of various diseases in various crops and also for solving different agricultural problems of various complexities and domains, by intelligent use of data mining and its tools such as cluster analysis, classification and prediction.

Keywords

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

1. Introduction

Agriculture is considered to be the oldest profession of mankind. Even in this modern era where there exits numerous kinds of businesses and professions, agriculture retains its supreme position in hundreds of the countries. The countries of the so called third world are still highly dependent on agriculture.

On the other hand, the present time is also considered as an era of knowledge and information. These days, there is virtually an explosion of information from every corner. Due to several vagaries, related to climate, pests and others, the condition of agriculture is getting in shambles in many parts of the world. 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 these days is to mine an agricultural knowledge base for discovering the hidden knowledge in these knowledge bases/data bases.

Before discussing application of cluster analysis in identifying the major diseases in soybean crop, 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 clusters, if comparison would have been 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, on the basis of different methods used for their categorization. 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.

Correct disease diagnosis is the prime requirement for recommending preventive or curative measures for effective disease management. We propose that, with the help of data mining and cluster analysis, in particular, we can even predict in advance, the future outbreak of any disease, based upon certain criterion, more accurately.

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 application of hierarchical agglomerative clustering methods for isolating/identification of the major factors responsible for the outbreak of diseases in the crop of soybean. 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/association. 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 SOYBEAN CROP DISEASES

We have chosen Soybean as an area of study because soy meal is world's most important vegetable protein feed source, accounting nearly 65% of world protein feed demand. Economically, soybean is the one of the most important bean in world.
Soybeans originated in Southeast Asia and were first domesticated by Chinese farmers around 1100 BC [6]. Soybeans were introduced to America in 1765 by Samuel Bowen, a sailor who had visited China. In 1932–33, the Ford Motor Company spent approximately $1,250,000 on soybean research. By 1935, every Ford car had soy involved in its manufacture. For example, soybean oil was used to paint the automobiles as well as fluid for shock absorbers [7].

For our study of identification of the major diseases in soybean crop using cluster analysis, we took the data base being provided by Machine Learning Repository, Center of Machine Learning and Intelligent System, University of California, Irvine [8].

Initially, there were 19 classes of diseases for soybean crop with 35 attributes. We then, removed 4 classes of diseases from the database as sufficient data were not available for those particular diseases thus leaving only 15 classes of diseases with us. As, we have decided to limited our study, only to the diseases of soybean, found in Indian subcontinent, we then cropped the data under study, only to reflect Indian soybean diseases.[10] This leaves with us, our dataset, containing 110 instances of data, spread over seven classes of diseases. For the purpose of clustering, these instances are treated as 110 different vectors of 35 dimensions each.

For creation of the clusters, we used Cluster 3.0 for windows and an associated program TreeView. The program Cluster was originally written by Michael Eisen while at Stanford University. Cluster and TreeView are programs that provide a computational and graphical environment for analyzing data from datasets. The program Cluster can organize and analyze the data in a number of different ways. TreeView allows the organized data to be visualized and browsed [9].

{We then re-order the data in random order so as to get an unbiased sample. Then, we used terms x1, x2…x289 respectively as object identifier. }

We then created clusters of above stated data set, using hierarchical agglomerative clustering method, taking Euclidean distance as measure of similarity and applying single linkage on it.

{After creating the clusters, we analyzed them using TreeView. The study of dendrogram, which graphically represents the clusters, revealed that objects x8, x12, x16, x14, x15, x135, x13,

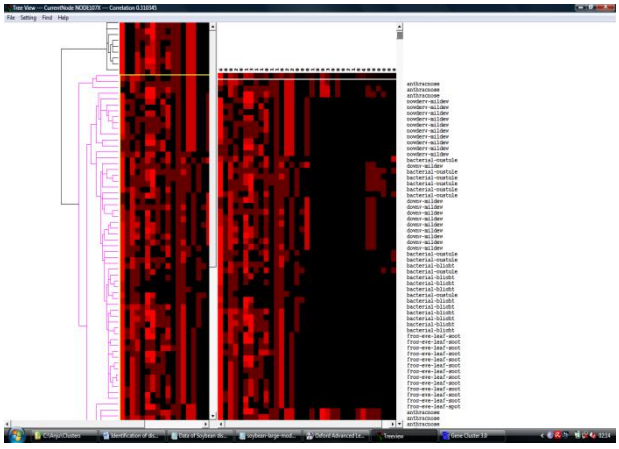
x7, x137, x136 are placed together in the given order. The colors, by which various attributes of the 35-dimenssional vector, describing the object, are indicative of the values of the attributes. }

Considering the value of the attributes, which is based on the shade of the colour in the dendrogram, and matching them with various conditions and criteria of soybean diseases, we found that the above listed data samples corresponds to the charcoal-rot, which is a soybean disease. The charcoal-rot is a root disease which spurs in periods of hot and dry conditions, which is evident from the dendrogram. Since it is a disease which affects the roots, some of the attributes may not involve in the final result, which may be vary from disease to disease. The matching conditions and criteria of soybean diseases, has been concerned with various literature on soybean diseases available over the internet and across many books [10][11].

We have also found that all the instances of the same disease are not placed in the same cluster. For example, the disease, Frog-eye-leaf spot has total 40 instances in all in the dataset, but it is split in two groups or clusters. A chunk of 13 records is placed after the cluster containing the records of the disease Bacterial-blight. Another chunk of 27 records is placed after the cluster containing the records of the disease Anthracnose.

Analyzing carefully the dendrogram, we can explain the reason of such a split. In the first cluster, representing the disease, Frog-eye-leaf spot, the following data is same for all instances, namely stem is normal, stem-cankers are absent, canker-lesion doesn't apply to them, external decay and fruit spots are absent while fruit pods are normal.

These characteristics are significantly different than those for another cluster representing the same disease. In this cluster the stem is abnormal, stem-cankers are present above second node, canker-lesion is apparent and it is brown in colour, external decay is also there and is firm and dry, fruit spots are coloured while fruit pods are diseased.



There is appearing a cluster that represents the disease Powdery-mildew, after the cluster representing the disease, Frog-eye-leaf spot. It is placed next to Frog-eye-leaf spot because both diseases share many common characteristics. As stem is normal, stem-canker is absent, external decay is absent, fruit pods are normal. Hence they are placed near Powdery-mildew and this points that the crop of late growing season ie. in months of July, August and September. Leaves are affected, but stem and fruits are normal in this cluster. So it divide the dataset for Frog-eye-leaf spot disease in two clusters. In one cluster, leaves are affected but not the stems while in other leaves as well as stems are abnormal. Also the fruit pods are also diseased. So they are placed in different clusters.

This is difference in the values of various attributes of the same diseases ie. Frog-eye-leaf spot and hence the instances are divided into two separate clusters which are at a distance as shown in the dendrogram.

The first cluster of Frog-eye-leaf spot disease dataset is placed before Powdery-mildew because Powdery-mildew is also a disease of leaves. Cool air temperature, low relative humidity are favourable conditions for it and so the good months for it is September and October. The leaves become abnormal and on the upper surface of leaf, there present leaf-mild [11]. This makes it different than Frog-eye-leaf spot in first cluster.

The second cluster of Frog-eye-leaf spot disease dataset is placed after Powdery-mildew because in the dataset of this cluster, leaves stem and fruit pods are affected. This is also near to the cluster representing the dataset for the disease Anthracnose, because there is only some mismatch as in Frog-eye-leaf spot disease mycelium is absent but in Anthracnose, it is present. In former the seeds are normal while seeds are abnormal in the case of later. Mold growth, seed discolour, shrivelling is absent in case of Frog-eye-leaf spot disease while in case of Anthracnose, they are present.

In the same way Bacterial-blight is also partitioned into two clusters. The period of time for Bacterial-blight in first cluster is July, August and September while in second cluster, it is June and July. The other difference is, in the case of first cluster leaf-malf is absent while in later case it is present.

There also appears the cluster representing the dataset for the disease Downy-mildew in between. The difference between the value of the attributes of Bacterial-blight and Downy-mildew is leaf-mild is absent in first case while it is on lower surface in later case.

Thus, using cluster analysis it is clear that in the month of June both the diseases Bacterial-blight and Downy-mildew appear. In the case of Downy-mildew the lower surface of the leaves is also affected but not in the case of Bacterial-blight.

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 upon the production of the food grains and pulses. 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] http://www.ncsoy.org/ABOUT-SOYBEANS/History-of-soybeans.aspx

[7] Schwarcz, Joseph A. (2004). The Fly in the Ointment: 70 Fascinating Commentaries on the Science of Everyday Life. Toronto: ECW Press. p. 193. ISBN 1-55022-621-5.

[8] http://archive.ics.uci.edu/ml/machine-learning-databases/ soybean /soybean-large.data

[9] Manual for Cluster 3.0 by Michael Eisen, Stanford University and updated by Michiel de Hoon, Human Genome Center, University of Tokyo.

[10] Diseases of Field Crops and Their Management-TS Thind, Daya Books

[11] www.ncsrp.org