IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

Role of Data Mining Techniques in Agriculture Improvement

Saurabh Sindhu[1], Divya Sindhu[2]

Lecturer, Department of Computer Science, CRM Jat College, Hisar, Haryana, India[1,2]

Abstract: Data mining is the process of discovering and extracting of interesting patterns and knowledge from large amounts of data. The field of agriculture has to deal with large amounts of data and processing and retrieval of significant data from this abundance of agricultural information is necessary to help the farmers. Therefore, appropriate methods and techniques are required for managing and organizing this data to increase the efficiency and agricultural productivity. The application of data mining methods and techniques to discover new insights or knowledge is a relatively a novel approach in agriculture. Data mining can help to process and convert this raw data into useful information for improving agriculture. In this paper, various data mining techniques used for processing of agricultural information/data such as k-means clustering, k-nearest neighbour, artificial neural networks, support vector machine, naive Bayesian classifier and fuzzy c-means are described. With the advancement of novel and appropriate data mining techniques, different types of agricultural problems will be addressed to improve crop productivity.

Keywords: Data mining, Agriculture productivity, k-means clustering, Artificial neural network, Naive Bayesian classifier, Association rule mining, Regression analysis.

I. INTRODUCTION

With an increasing population and a commensurate need for increasing agricultural production, there is an urgent need to improve management of agricultural resources [1]. Nearly two-thirds of population in India directly depends on agriculture for its livelihood and therefore, agriculture is the backbone of the Indian economy. The productivity of agriculture is very low because only one-third of the cropped part is irrigated [2, 3]. So as the demand of food is increasing, the farmers, agricultural scientists and government are trying to put an extra effort by implementing techniques for more food production. But still today, farmers are performing agriculture-related tasks manually and a very few farmers are using the new methods, tools and techniques of farming for better agriculture production. Moreover, in traditional crop field management, uniform input application not only consider the concept of spatial and temporal variability within a crop field, but also results in environmental pollution and reduction of farm profits. The need of site-specific management or precision agriculture has been advocated by researchers, producers and farmers in the worldwide. Advanced information technology that can provide quick and cost-effective ways to identify spatial variability within crop fields is the basis of precision agriculture. Moreover, remote sensing technologies have advanced rapidly in recent years and have become effective tools for site-specific management in crop protection and production.

Raw data in agriculture is very diverse. It is necessary to collect and store it in an organized form and integrate it for the creation of agricultural information system. Therefore, there is a need of utilization of information and communications technologies, which will enable the extraction of significant data from agriculture in an effort to obtain knowledge and trends. It will also help in the elimination of manual tasks. Easier data extraction directly from electronic sources and its transfer to secure electronic system of documentation, will reduce the production cost, higher yield and higher market price. Data mining technique will provide information about crops and enable agricultural enterprises to predict trends about customer's conditions or their behaviour. It analyzes the data from different perspectives and helps in finding relationships in seemingly unrelated data. The computational needs of agriculture data and how data mining techniques can be used as a tool for knowledge management in agriculture should be considered by researchers. Data warehouses can be prepared to hold agriculture data, which makes transaction management, information retrieval and data analysis much easier.

Data mining represents a set of specific methods and algorithms aimed solely at extracting and patterns patterns from raw data. It helps in the process of discovering previously unknown and potentially interesting patterns in large datasets [4]. Data mining, which is also termed as knowledge discovery, is the process of analyzing data from different perspectives and summarizing it into valuable information for future use [5, 6]. This 'mined' information produced from data mining can be used for variety of purposes like research, future forecasting or prediction, classification etc. in agriculture. Analysis of data in effective way requires understanding of appropriate techniques of data mining. The objective of this paper is to describe different data mining techniques in perspective of agriculture domain.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

II. METHODS USED: APPLICATION OF DIFFERENT ALGORITHMS

Data mining tasks can be classified into two categories: Descriptive data mining and Predictive data mining. Descriptive data mining tasks characterize the general properties of the data in the database while predictive data mining is used to predict explicit values based on patterns determined from known results. Prediction involves using some variables or fields in the database to predict unknown or future crop, weather forecasting, use of pesticides and fertilizers to be used and revenue to be generated. Different techniques and algorithms used in data mining for improving agriculture productivity are described in this section.

1. Clustering

Clustering is an unsupervised learning technique that takes unlabeled data points (data records) and classifies them into different groups or clusters. This is done in such a way that points assigned to the same cluster have high similarity, while the similarity between points assigned to different clusters is low [7]. In clustering, the focus is on finding a partition of data records into clusters such that the points within each cluster are close to one another. It can also be defined as a process which partitions a set of data (or objects) into a set of meaningful sub-classes called clusters [8].

1.1. k-means approach: The k-means is a data mining technique for clustering [9, 10]. The aim is to find a partition of the set in which similar data are grouped in the same cluster given a set of data with unknown classification. Samples that are close to each other are considered similar and the measure of similarities between data samples is calculated using a suitable distance. The parameter k in the k-means algorithm plays an important role as it specifies the number of clusters in which the data must be partitioned. The center of the cluster can be considered as the representative of the cluster because the center is quite close to all samples in the cluster. It follows that a cluster contains similar data if all its samples are closer to its center and not to the center of some other cluster. Therefore, the k-means algorithm moves the corresponding data samples from their original cluster to the new cluster, when samples belonging to a cluster are closer to the center of a different cluster. Figure 1 shows an optimal partition in clusters of a set of points in a cartesian space.

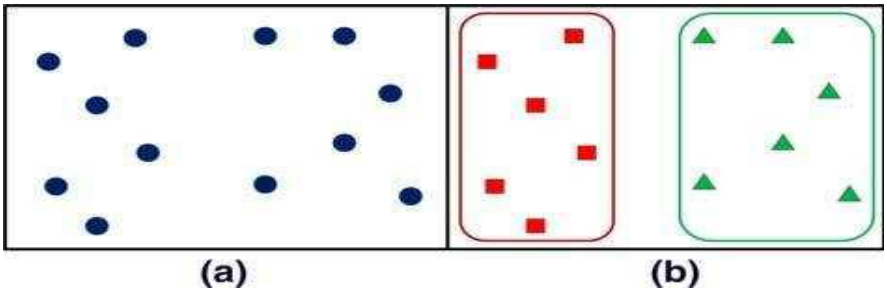

Figure 1. An optimal partition in clusters of a set of points in a cartesian space: (a) the points are not assigned yet to any cluster; (b) points belonging to the same cluster are marked using the same symbol.

The k-means algorithm is an algorithm for local optimization, because it identifies a sequence of partitions in clusters which have strict decreasing function values. Hence, the k-means algorithm is able to find only one of the local minima of the function f, that may or may not correspond to the global minimum. For this reason, the k-means algorithm is usually performed several times using different initial partitions. The partition corresponding to the smallest value of the function f is considered to be the optimal solution. The k-means algorithm belongs to the category of expectation maximization (EM) algorithms that are elegant and powerful methods for finding maximum likelihood solutions for models with latent variables [11].

There are some disadvantages in using the k-means algorithm. One of the disadvantages could be the choice of the parameter k. In some applications, this choice can be trivial, if the samples have to be divided in a predetermined number of categories, such as ‘‘good’’ and ‘‘bad’’ samples, then in this case k = 2. In the majority of cases, however, the parameter k is unknown apriori. The computational cost of the algorithm is another issue that needs attention. Most of the cost is due to the computation of distances and new centers. The standard k-means algorithm computes new centers every time a sample migrates from one cluster to another. There is a variant of the standard algorithm that computes new centers only when all samples in the set of data have been checked and eventually moved. This variant is referred to as the h-means algorithm. The two algorithms are sometimes combined together in which the h-means algorithm can be initially used to identify a partition close to the optimal one that is then used by the k-means algorithm as a starting partition. Over the years, other variants of the basic k-means algorithm have been proposed. The most popular implementation of this algorithm is the Lloyd algorithm [12].

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

1.2. Bi-clustering: Bi-clustering of a set of data, is actually a technique for classification that exploits the information from a training set. A set of data is basically formed by samples, which are represented by a sequence of features that are considered to be relevant for the representation of the samples. Instead of considering samples only, bi-clustering aims at finding simultaneous classifications of samples and of their features. Moreover, if a training set is known, a bi-clustering can be constructed by exploiting this training set. The corresponding partition in bi-clusters is able to associate subgroups of samples to subgroups of features, so that the features causing the classification of the training set are revealed. This information can then be exploited for performing classification of samples which do not belong to the training set.

1.3. Fuzzy clustering: Studies were conducted by using fuzzy clustering in detection of leaf spots in cucumber crop [13, 14]. Spots on leaves gave an indication of plant diseases, which were examined manually and were then subjected to expert advice. Then the experts declared the disease after proper investigation. Scientists proposed a segmentation technique for identifying leaf batches in cucumber crop using fuzzy clustering algorithm. The first step of image analysis and pattern recognition is the segmentation of image [15]. Segmentation is very critical and inevitable component of image analysis and pattern recognition. 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.

2. Classification

Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. It is a process in which a model learns to predict a class label from a set of training data which can then be used to predict discrete class labels on new samples. A classification problem arises when an object needs to be assigned to a predefined class or group according to its characteristics. A classification task is also referred to as a supervised learning task since the classes or groups are defined before hand and can be used to steer the learning process.

2.1. k-nearest neighbour: k-nearest neighbour classifiers (KNN) classify a data instance by considering only the k most similar data instances in the training set [16]. The class label is then assigned according to the class of the majority of the k-nearest neighbours. A training set is known and it is used to classify samples of unknown classification. The basic assumption in the k-NN algorithm is that similar samples should have similar classification. As in the k-means approach, the similarities between samples are measured using suitable distance functions. A sketch of the k-NN algorithm is given in Fig. 2.

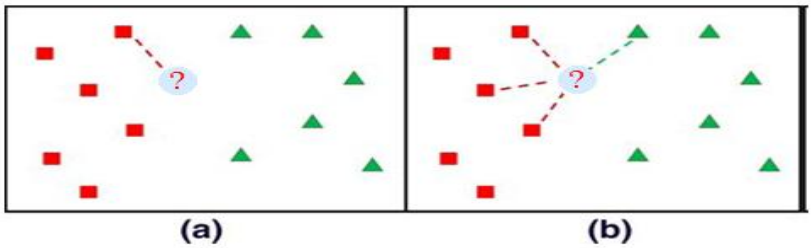

Figure 2. The point marked by the symbol ? is classified according to the classification of its nearest neighbors: (a) k = 1 and the unknown point is classified as belonging to the class marked by squares; (b) k = 4 and the unknown point is classified as belonging to the class marked by squares as well.

2.2. Support vector machines (SVMs): These machines are binary classifiers that are able to classify data samples in two disjoint classes [18].

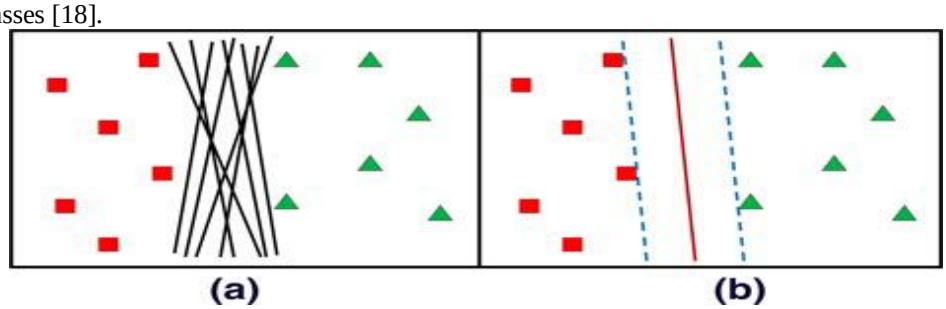

Figure 3. Points in a Cartesian system are separated on the basis of their features and are assigned to two different classes: (a) possible separating hyperplanes that do not maximize the margin between the two classes; (b) the separating hyperplane found by SVMs, providing the maximum margin.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

In this technique, two considered classes are linearly separable. In such a case, there exists a hyperplane which is able to separate all samples in two classes. Actually, in most of the cases, more than one hyperplane satisfying this condition exists and one of them is chosen as classifier on the basis of the margin it creates between the two classes. Intuitively, the larger is the margin, the less are the possibilities of misclassifications. Figure 3 shows points in a cartesian system to be classified in two different classes.

2.3. Decision trees: A decision tree is a flowchart-like structure with two types of components: (i) Leaf nodes that assign class labels to observations, (ii) Internal nodes that specify tests on individual attributes and each branch represents an outcome of the test [19]. The tree classifies observations in a top-down manner, starting from the root and it moves down according to the outcomes of the tests at the internal nodes until a leaf node has been reached and a class label has been assigned. The tree is then constructed by means of recursive partitioning until the current leaf nodes contain only instances of a single class or until no test offers any improvement. This tree growing strategy often results in a complex tree with many internal nodes that overfits the data because most of the real-life data sets are noisy and in most of the cases, the attributes have limited predictive power.

2.4. Bayesian network: Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. This model has several advantages for data analysis when used in conjunction with statistical techniques [20]. Firstly, the model encodes dependencies among all variables, therefore, it can readily handle situations where some data entries are missing. Secondly, a Bayesian network can be used to learn causal relationships and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Thirdly, it is an ideal representation for combining prior knowledge and data because the model has both a causal and probabilistic semantics. Moreover, Bayesian statistical methods in conjunction with Bayesian networks offer an feasible and efficient approach for avoiding the over fitting of data. It is a powerful tool for dealing the uncertainties, which widely occur in agriculture data sets.

2.5. Artificial neural networks: Artificial neural networks (ANNs) are systems inspired by the research on human brain [21]. In these networks, each node represents a neuron and each link represents the interaction between two neurons. Each neuron performs very simple tasks, while the network, representing of the work of all its neurons, is able to perform more complex tasks (Fig. 4). The ability of a neural network to perform a given task depends on the structure of the network. The most commonly used ANNs is the multilayer perceptron, in which neurons are organized in layers. The input layer contains neurons that receive the input signal which is then fed to the network. The neurons on the output layer are active and the result provided by them is considered as the output generated by the network. There are also hidden layers between the input and the output layers. Each neuron can receive input signals from the neurons belonging to the previous layer and it sends its output to neurons belonging to the successive layer. The organization of neurons in layers and their interconnections define the structure of the multilayer perceptron.

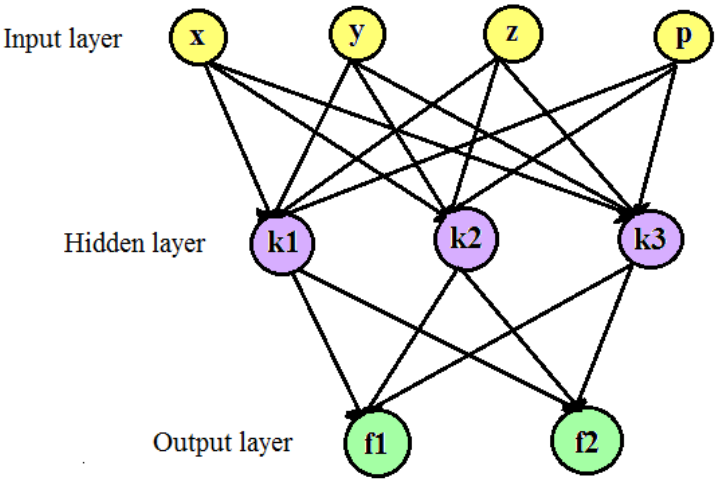

Figure 4. The scheme of a multilayer perceptron. In this figure, there is only one hidden layer which contains three neurons.

3. Genetic Algorithm

It is a method of optimization and research technique which uses techniques based on evolution, specifically inheritance, mutation and selection [22]. Genetic algorithm at the same time processes a set (population) of potential

solutions (individual) for a given problem. Algorithm begins with a set of solutions called sub-population and it creates certain number of random solutions. Suitability of those solutions based on criteria of performance is evaluated and used to select solutions making a new, better subset of potential solutions [23]. Bad solutions are disregarded and good solutions are kept. Good solutions are then hybridized and the whole process is repeated. In the end, similar to the process of natural selection, only the best solutions are chosen and combined with each other with the purpose of getting one universal solution from the set of solutions, similar to the process of organism population evolution. Genetic algorithms are used in data mining to formulate hypotheses about variable dependencies, in the shape of association rules or some other internal formalism [24]. Careful selection of structure and parameters of the genetic algorithm can secure a significant chance to come up with globally optimal solution after an acceptable number of iterations.

4. Association rule mining

Association rule mining is the technique of discovering association rules used to search unseen or desired pattern among the vast amount of agricultural data [25]. In this method, the focus is on finding relationships between the different items in a transactional database. Association rules are used to find out elements that co-occur repeatedly within a dataset consisting of many independent selections of elements (such as purchasing transactions) and to discover rules. An application of the association rules mining is the market basket analysis, customer segmentation, catalog design, store layout and telecommunication alarm prediction [26]. Apriori algorithm is one of the association rule mining algorithm.

5. Regression analysis

Regression analysis is a statistical tool that uses the relation between two or more quantitative variables so that one variable (dependent variable) can be predicted from the other variables (independent variables). Regression is a method of finding correlation between different metric variables, fields or datasets. It is learning of a function that analyses and provides a data item into real valued prediction figure. Strong or weak relationship between the variables is also calculated based on certain assumptions. The strength of the system or at what level the considered model is fitted can be done by regression analysis. Many regression analysis models are available including simple linear, multiple linear, curvilinear and multiple curvilinear regression models. Basically, regression takes a numerical dataset and develops a mathematical formula that fits the data. Advanced techniques are available such as multiple regression that allow the use of more than one input variable for the fitting of more complex models such as a quadratic equation. Multiple linear regression (MLR) is the method used to model the linear relationship between a dependent variable and one or more independent variable(s). The dependent variable is sometimes termed as predicting i.e. rainfall and independent variables are called predictors i.e., year, area of sowing and production.

III. APPLICATION OF DATA MINING IN AGRICULTURE

Data mining techniques have already been successfully applied in various fields including engineering, medicine, education, marketing etc. In agriculture, farmers have many queries regarding the kind of soil and climatic conditions for cultivation of a specific crop and the timelines corresponding with each activity related to agriculture. Recently, a number of studies have been carried out on the application of data mining techniques for agricultural data sets.

1. Classification of soils and prediction of soil fertility

The improved clustering algorithm is a good method for comprehensive evaluation of soil fertility. The k-means algorithm is used for soil classifications using GPS-based technologies [27], classification of plant, soil and residue regions of interest by color images [28]. Decision tree approach technique is also used in the prediction of soil fertility [29]. In addition, naive Bayesian classification technique is used to classify soils that analyze large soil profile experimental datasets [30]. A range of spectral reflectance patterns in the visible and infrared range were examined by deploying remote sensing for detection of plant stress, particularly nutrient deficiency [31]. This approach can potentially lower operating cost of fertilization and may minimize loss of crop productivity.

2. Relationship between sprays and fruit defects

Fruit defects are caused for a multitude of reasons including physical injury or the damage caused by birds, pathogens and insects. These defects may be recorded manually or through computer vision (detecting surface defects when grading fruit). Spray diaries are a legal requirement in many countries and record the date of spray and the product name. It is known that spraying affects different fruit defects for different fruits. Fungicidal sprays are often used to prevent rots from being expressed on fruit. It is also known that some sprays can cause russeting on apples. Grading of apples is done before marketing using k-means clustering [32]. Currently some efforts have been made with regard to the use of data mining in horticulture [33].

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

3. Forcasting of weather and rainfall

For simulation of daily precipitations and other weather variables, k-nearest neighbor approach technique can be applied [34]. It also helps in the forecasting of climate and in the estimation of soil water parameters which are referred to as lower limit of plant water availability (LL), drained upper limit (DUL) and plant extractable soil water (PEWS) [35]. With the advancement of satellite remote sensing technologies, the forest and agricultural land observations have become very convenient. These sensors produce potentially useful data in enormous quantity on daily basis. This quantity of data also presents a challenge of data interpretation and classification to the researchers. Scientists have been widely using techniques such as k-means, k-nearest neighbour and artificial neural network for classification of remotely sensed images.

Somvanshi et al. [36] designed the model for the prediction of rainfall using artificial neural networks and Box-Jenkins methodology. Other applications of artificial neural networks in hydrology are forecasting daily water hassle and flow forecasting. Sawaitul et al. [37] focused the information about weather and the recorded parameters were used to forecast weather. If there is a change in any one of the recorded parameters like wind speed, wind direction, temperature, rainfall and humidity, then the upcoming climatic conditions can be predicted using artificial neural networks and back propagation techniques. Farmers are being issued information from the state agricultural universities and government kissan portals through e-mails or SMS about the weather conditions, sowing of crops and prevalence/prevention of pathogens and pests in different geographical regions and agro-ecosystems.

4. Prediction of crop yield

For estimation and analysis of the crop yield, k-means clustering is used [38]. Jagielska et al. [39] described applications of data mining to agricultural related areas in relation to yield prediction. In the past, yield prediction was achieved by considering farmer's experience on particular field, crop and climate condition. The additional information about data like probability in probability theory, grade of membership in fuzzy set theory could also be discussed. The neural network is used in prediction of flowering and maturity dates of soybean [40] and in forecasting of water resources variables [41].

Veenadhari [42] studied the influence of climatic factors on major kharif and rabi crops production in Bhopal district of Madhya Pradesh state. The findings showed that the productivity of soybean crop was mostly influenced by comparative humidity followed by temperature and rainfall by using the decision tree analysis technique. The same technique showed that the productivity of paddy crop was mostly inclined by rainfall followed by comparative evaporation and humidity. For wheat crop, the analysis revealed that the productivity is mostly influenced by temperature followed by relative humidity and rainfall. The result of decision tree was confirmed from Bayesian classification. The rules formed from the decision tree were useful for identifying the conditions required for high crop productivity.

5. Planning of grain, oil seed or cash crops for cultivation

Genetic algorithms can be used for the process of decision making with the purpose of finding appropriate crops to grow, which will be highly profitable to our farmers [43]. Support vector machine technique may be used in area of the crop classification [44] and in the analysis of the climate change scenarios [45]. It can also be used for detecting weed and nitrogen stress in corn [46].

6. Control of weeds among the crop plants in the field

Weeds are unwanted useless plants that compete with crop plants for space, nutrients, water, sunlight and other elements. These weed plants reduce the biomass and yield of the main crop. Thus, weeds pose a serious constraint to agricultural production and usually result in average ~ 12% losses of the world's agricultural output. Therefore, weed control is indispensable in every crop production system. The losses due to weed growth include interference with cultivation of crops, loss of biodiversity, loss of potentially productive lands, loss of grazing areas and livestock production, choking of navigational and irrigation canals and reduction of available water in water bodies. They decrease quantity and quality of produce/food, fibre, oil, forage/fodder and animal products (meat and milk), and cause health hazards to humans and animals. Weeds force the use of large amounts of human labour and technology to prevent greater crop losses.

Goel et al. [47] reported a strong correlation between the digital information, e.g., spectral data of the aerial image and soybean crop physiological parameters such as chlorophyll fluorescence, leaf greenness, leaf area index, photosynthesis rate and plant height. They used multi-spectral (24 wave band with a range of 475.12 nm to 910.01 nm) airborne optical remote sensing technique for detecting weed infestation in site-specific managed field crops. Their results indicated that the wave band centered on 675.98 and 685.17 nm in the red region and 743.93 to 830.43 nm in the NIR region, had good potential for weed classification in a maize field. The Schiffes multiple range tests provided a p-value that was less than 0.05 to support their findings. Tellaeche et al. [48] purposed an approach for the detection of weeds in agriculture and summarized an automatic computer vision system for the detection and differential spraying of Avena

sterilis, a toxic weed found in cereal crops. So, a hybrid decision making system based on the Bayesian and Fuzzy k-means classifiers has been designed where the apriori probability required by the Bayes framework is supplied by the Fuzzy k-means.

7. Optimizing the use of pesticide by data mining

Global crop yields are reduced by 20 to 40% annually due to attack of plant pests and pathogens. For the control of pests and phytopathogens in agriculture, farmers have mostly relied on the application of synthetic pesticides and the global pesticide market is presently growing at a rate of 3.6% per year [49]. However, indiscriminate use of chemical pesticides to control the pathogens/insects has generated several problems including resistance to insecticides/fungicides, outbreak of secondary pests as well as safety risks for humans and domestic animals. Moreover, the long persistence of applied pesticides in soil leads to contamination of ground water and soil, and the residual toxic chemicals may enter in the food chain. Hence, excessive use of pesticides is harming the farmers with adverse financial, environmental and social impacts.

Recent studies showed that attempts of cotton crop yield maximization through pro-pesticide state policies have led to a dangerously high pesticide use. Coarse estimates of the cotton pest scouting data recorded stands at around 1.5 million records and growing. The primary agro-met data recorded has never been digitized, integrated or standardized to give a complete picture and hence, cannot support decision making. These studies have reported a negative correlation between pesticide use and crop yield. By data mining, using the cotton pest scouting data along with the meteorological recordings, it was shown that how pesticide use can be optimized (reduced). Clustering of data revealed interesting patterns of farmer practices along with pesticide use dynamics and hence help in identifying the reasons for this pesticide abuse [50]. Creating a novel Pilot Agriculture Extension Data Warehouse followed by analysis through querying and data mining some interesting discoveries were made, such as pesticides sprayed at the wrong time, wrong pesticides used for the right reasons and temporal relationship between pesticide usage and day of the week [51].

8. Sorting apples by watercores

Before going to market, apples are checked and the apples showing some defects are removed. However, there are also invisible defects that can spoil the apple flavor and look. An example of invisible defect is the watercore. This is an internal apple disorder that can affect the longevity of the fruit. Apples with slight or mild watercores are sweeter, but apples with moderate to severe degree of watercore cannot be stored for any length of time. Moreover, a few fruits with severe watercore could spoil a whole batch of apples. For this reason, a computational system is under study which takes X-ray photographs of the fruit while they run on conveyor belts [52] and which is also able to analyse (by data mining techniques) the taken pictures and estimate the probability that the fruit contains watercores [53]. Neural network is also applied for discrimination between good and bad apples.

9. Prediction of problematic wine fermentations

Wine is widely produced from grapes all around the world. The fermentation process of the wine is very important, because it can impact the productivity of wine-related industries and also the quality of wine. If the fermentation defect could be categorized and predicted at the early stages of the process, it could be altered in order to guarantee a regular and smooth fermentation. Fermentations are nowadays studied by using different techniques such as the k-means algorithm [54] and a technique for classification based on the concept of bi-clustering [55]. Urtubia et al. [56] demonstrated that the prediction of wine fermentation problems can be performed by using a k-means approach. Knowing in advance that the wine fermentation process could get jammed or be slowed, it can help the enologist to correct it and ensure a good fermentation process. Moreover, taste sensors are used to obtain data from the fermentation process to be classified using ANNs [57].

10. Prediction of metabolizable energy of poultry feed using group method of data handling-type neural network

A group method of data handling-type neural network (GMDH-type network) with an evolutionary method of genetic algorithm was used to predict the metabolizable energy of feather meal and poultry meal based on their protein, fat and ash content [58]. Published data samples were used to train a GMDH-type network model. It is also reported that the GMDH-type network may be used to accurately estimate the poultry performance from their dietary nutrients such as dietary metabolizable energy, protein and amino acids [59].

11. Detection of diseases from sounds issued by animals by neural networks

The detection of animal's diseases in farms can impact positively the productivity of the farm, because sick animals can cause contaminations. Moreover, the early detection of the diseases can allow the farmer to cure the animal as soon as the disease appears. Coughing, in human and animals, is associated with the sudden expulsion of air and it is typically accompanied with a sound, whose changes may reflect the presence of diseases. The sound provided by pigs due to coughing can be used to monitor possible health problems. An expert could analyze whether the cough of a pig signals

the presence of a potential disease and eventually check the health of the pig. Systems for the automatic control of the pig houses are useful to avoid the infection of humans because of the presence of contagious diseases. Therefore, considerable efforts have been undertaken to develop and apply sensing techniques for diagnosis of diseases in pig farms. The early detection of animal diseases can bring on the consumer's table better meat, by reducing, for instance, the residuals of antibiotics.
A neural network approach for cough recognition is described [60]. The training set is obtained by experimental observations, where the sounds produced by pigs are recorded and where each record is labeled by an expert in different ways. A metal construction has been built in order to perform the experiments, where pig sounds are recorded. The construction is covered with transparent plastic material for controlling the environment around the animal. The time signal of these sounds is analyzed mathematically and transformed in a vector formed by 64 real numbers. The vectors are normalized before the use, because their components can variate significantly even when comparing two vectors from the same class. These variations are mainly due to the distance and direction between the pigs and the microphone. A neural network is trained using the training set obtained during the experiments. The used network is a multilayer perceptron with one hidden layer of hyperbolic tangent neurons, while the output layer consists of logistic neurons. The network is firstly trained to discriminate between coughs and metal clanging, and it is able to reach percentages of correct recognition greater than 90%. Successively, the network is trained in order to distinguish among four sounds: coughs, metal clanging, grunting and background noise.

12. Growth of sheep from genes polymorphism using artificial intelligence
Polymerase chain reaction-single strand conformation polymorphism (PCR-SSCP) method was used to determine the growth hormone (GH), leptin, calpain and calpastatin polymorphism in Iranian Baluchi male sheep. An artificial neural network (ANN) model was developed to describe average daily gain (ADG) in lambs from input parameters of GH, leptin, calpain and calpastatin polymorphism, birth weight and birth type. The results revealed that the ANN-model is an appropriate tool to recognize the patterns of data to predict lamb growth in terms of ADG given specific genes polymorphism, birth weight, and birth type. The platform of PCR-SSCP approach and ANN-based model analyses may be used in molecular marker-assisted selection and breeding programs to design a scheme in enhancing the efficacy of sheep production [61].

13. Analyzing chicken performance data by neural network models
A platform of artificial neural network-based models with sensitivity analysis and optimization algorithms was used successfully to integrate published data on the responses of broiler chickens to threonine. Analyses of the artificial neural network models for weight gain and feed efficiency from a compiled data set suggested that the dietary protein concentration was more important than the threonine concentration. The results revealed that a diet containing 18.69% protein and 0.73% threonine may lead to producing optimal weight gain, whereas the optimal feed efficiency may be achieved with a diet containing 18.71% protein and 0.75% threonine [62].

14. Detection of meat and bone meal by support vector machines
Since the emergence of the mad cow crisis in Europe and all its socio-economic consequences, European Union regulatory agencies have established many legal measures to assure the safety and the quality of feedstuffs for animals. One of the most important decisions is to ban meat and bone meal in feedstuffs destined to farm animals which are kept, fattened or bred for food production. Controls are needed to verify whether meat and bone meal is used for accidental contamination or against the law. Therefore, the effective enforcement of this regulation requires accurate and efficient analytical methods capable of analyzing thousands of samples per year.
Near-infrared microscopy method has been developed, which works well in discriminating the different ingredients found in compound feeds. Each particle in the feedstuffs is evaluated on the basis on its chemical properties rather than its appearance. Recently, new methods have been developed that combine the advantages of spectroscopic and microscopic methods along with much faster sample analysis. The goal is to gather spectral and spatial data simultaneously by recording sequential images of a predefined sample. The set of data obtained using this method (a collection of spectra) can be used for training a SVM with the aim of defining a classifier able to discriminate between vegetable and meat and bone meal [63].

Spectra coming from 26 pure animal meals and spectra coming from 59 pure vegetable meals are used for creating a training set. The analyzed animal and vegetable materials are selected to span the diversity of materials which are mainly used for the formulation of compound feeds. In total, more than 267,000 spectra were collected from pure animal and vegetable meals. In this application, samples belonging to two classes only have to be discriminated and therefore only one SVM is needed. Different kernels have been tested using the obtained set of data and the results show that the best choice is the Gaussian kernel, which provides accurate predictions. In this particular area, it is very important that all samples are classified with a good precision. Human analysis can be included in the process for verifying whether there are false detections of meat and bone meal, but this would require additional expenses.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

IV. CONCLUSION

Recent technologies are able to provide a lot of information on agricultural-related activities, which can then be analyzed in order to find important information. There is large amount of data in agriculture that is currently available from many resources and many applications of data mining techniques are recently being used in agriculture. The application of data mining techniques in agriculture is relatively a novel research field. In a near future, more sophisticated techniques can be developed to address complex problems in agriculture-related fields and hence may provide better results. A lot of work has to be done on this emerging and interesting research field involving multidisciplinary approach consisting of mathematicians and computer scientists to help agronomists and farmers in finding solutions to the complex problems to impove agriculture productivity leading to sustainable agriculture.

REFERENCES

[1] Wezel, A., Casagrande, M., Celette, F., Vian, J.F., Ferrer, A. and Peigne, J., "Agroecological practices for sustainable agriculture. A review", Agronomy and Sustainable Development, Vol. 34, pp. 1-20, 2014.
[2] Shiklomanov, I.A., "Appraisal and assessment of world water resources", Water International, Vol. 25, Issue 1, pp. 11–32, 2000.
[3] Siebert, S., Burke, J., Faures, J.M., Frenken, K., Hoogeveen, J., Doll, P. and Portmann, F.T., "Groundwater use for irrigation – a global inventory", Hydrology and Earth System Science, Vol. 14, pp. 1863–1880, 2010. doi:10.5194/hess-14-1863-2010
[4] Fayadd, U., Piatesky–Shapiro, G. and Smyth, P., "Data mining to knowledge discovery in databases", pp. 50-67, 1996.
[5] Naik, R. and Deepika, N., "Data mining system and applications: A study", International Journal of Computer Science and Mobile Computing, Vol. 5, Issue 12, pp. 103-110, 2016.
[6] Sindhu, S. and Sindhu, D., "Data mining and gene expression analysis in bioinformatics", International Journal of Computer Science and Mobile Computing, Vol. 6, Issue 5, 72-83.
[7] Han, J., Kamber, M. and Pei, J., "Data mining: Concepts and techniques", 3rd edition, The Morgan Kaufmann Series in Data Management Systems, USA, 2006.
[8] Lee, R.C.T., "Cluster analysis and its applications", In: Advances in Information Systems Science, pp. 580-584, 1981.
[9] Hartigan, J., "Clustering algorithms", John Wiles & Sons, New York, 1975.
[10] Sindhu, S. and Singh, S., "Clustering algorithms: Mean shift and K-means algorithm", International Journal of Technical Research, Vol. 3, issue 2, pp. 1-6, 2014.
[11] Dempster, A.P., Laird, N.M. and Rubin, R.D., "Maximum likelihood from incomplete data via the EM algorithm", Journal of Research of Statistical Society, Vol. B 39, Issue 1, pp. 1–38, 1977.
[12] Lloyd, S.P., "Least squares quantization in PCM", IEEE Trans Information Theory, Vol. 28, Issue 2, pp. 129–137, 1982.
[13] "Fuzzy Clustering" in http://en.wikipedia.org/wiki/Fuzzy_clustering
[14] Helly, M. El., Onsi, H., Rafea, A., El-Gamma, S., "Detecting leaf spots in cucumber crop using fuzzy clustering algorithm".
[15] Sindhu, S. and Singh, S., "Image segmentation in various domains using two phase clustering approach", International Journal of Emerging Trends and Technology in Computer Science, Vol. 3, issue 4, pp. 173-176, 2014.
[16] Cover, T.M. and Hart, P.E., "Nearest neighbor pattern classification", IEEE Trans Information Theory, Vol. 13, issue, 1, pp. 21–27, 1967.
[17] Hart, P.E., "The condensed nearest neighbor rule", IEEE Trans Information Theory, Vol. 14, pp. 515–516, 1968.
[18] Burges, C.J.C., "A tutorial on support vector machines for pattern recognition", Data Mining and Knowledge Discovery, Vol. 2, Issue 2, pp. 955–974, 1998.
[19] Quinlan, J.R., "C4.5 programs for machine learning", Morgan Kaufmann, 1993.
[20] Pollino, C.A., White, A.K. and Hart, B.T., "Examination of conflicts and improved strategies for the management of an endangered Eucalypt species using Bayesian networks", Ecological Modelling, Vol. 201, pp. 37–59, 2007
[21] Nurnberger, A., Pedrycz, W. and Kruse, R., "Neural network approaches", In: Klosgen, W. and Zytkow, J.M. (eds), Handbook of data mining and knowledge discovery. Oxford University Press, 2002.
[22] Agarwal, R., "Genetic algorithm in data mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, Issue 9, pp. 631-634, 2015.
[23] Huang, Y., Lan, Y., Thomson, S.J., Fang, A., Hoffmann, W.C. and Lacey, R.E., "Development of soft computing and applications in agricultural and biological engineering", Computers and Electronics in Agriculture, Vol. 71, pp. 107–127, 2010.
[24] Kudjao, P.K., Ocquaye, E.N.N. and Ametepe, W., "Reveiew of genetic algorithm and application in software testing", International Journal of Computer Applications, Vol. 160, Issue 2, pp. 1-6, 2017.
[25] Agrawal, R., Imieliński, T. and Swami, A., "Mining association rules between sets of items in large databases", ACM Sigmod Record, Vol. 22, Issue 2, pp. 207-216, 1993.
[26] Zaki, M.J., "Parallel and distributed association mining: A survey", IEEE Concurrency, Vol. 7, Issue 4, pp. 14-25, 1999.
[27] Verheyen, K., Adriaens, D., Hermy, M. and Deckers, S., "High-resolution continuous soil classification using morphological soil profile descriptions", Geoderma, Vol. 101, Issue 3, pp. 31-48, 2001.
[28] Meyer, G.E., Camargo Neto, J., Jones, D.D. and Hindman, T.W., "Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images", Computers and electronics in agriculture, Vol. 42, Issue 3, pp. 161-180, 2004.
[29] Gholap, J., "Performance tuning of j48 algorithm for prediction of soil fertility", Asian Journal of Computer Science and Information Technology, Vol. 2, Issue 8, pp. 251– 252, 2012.
[30] Bhargavi, P and Jyothi, S., "Applying naive Bayes data mining technique for classification of agricultural land soils", International Journal of Computer Science and Network Security, Vol. 9, Issue 8, pp. 117-122, 2009.
[31] Mee, C.Y., Balasundram, S.K. and Hanif, A.H.M., "Detecting and monitoring plant nutrient stress using remote sensing approaches", Asian Journal of Plant Sciences Vol. 16, pp. 1-8, 2017.
[32] Leemans, V. and Destain, M. F., "A real-time grading method of apples based on features extracted from defects", Journal of Food Engineering, Vol. 61, Issue 1, pp. 83-89, 2004.
[33] Hill, M.G., Connolly, P.G., Reutemann, P. and Fletcher, D., "The use of data mining to assist crop protection decisions on kiwifruit in New Zealand", Computers and Electronics in Agriculture, Vol. 108, pp. 250–257, 2014. doi:10.1016/j.compag.2014.08.011.
[34] Rajagopalan, B. and Lall, U., "A k–nearest-neighbor simulator for daily precipitation and other weather variables", Water Resources Research, Vol. 35, Issue 10, pp. 3089-3101, 1999.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

[35] Jagtap, S.S., Lall, U., Jones, J.W., Gijsman, A.J. and Ritchie, J.T., "Dynamic nearest-neighbor method for estimating soil water parameters", Trans ASAE Vol. 47, Issue 5, pp. 1437–1444, 2004
[36] Somvanshi, V.K., Pandey, O.P., Agrawal, P.K., Kalanker, N.V., Prakash, M.R. and Chand, R., "Modeling and prediction of rainfall using artificial neural network and ARIMA Techniues", Journal of Indian Geophysics Union, Vol. 10, Issue 2, pp. 141-151, 2006.
[37] Sawaitul, S.D., Wagh, K.P. and Chatur, P.N., "Classification and prediction of future weather by using back propagation algorithm- An approach", International Journal of Emerging Technology and Advanced Engineering, Vol. 2, Issue 1, pp. 110-113, 2012.
[38] Ramesh, D. and Vardhan, V.B., "Data mining techniques and applications to agricultural yield data. International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 9, pp. 26-31, 2013.
[39] Jagielska, I., Mattehews, C. and Whitfort, T., "An investigation into the application of neural networks, fuzzy logic, genetic algorithms, and rough sets to automated knowledge acquisition for classification problems", Neurocomputing, Vol. 24, pp. 37-54, 1999.
[40] Elizondo, D.A., McClendon, R.W. and Hoogenboom, G., "Neural network models for predicting flowering and physiological maturity of soybean", Transactions of the ASAE (USA), 1994.
[41] Maier, H.R. and Dandy, G.C., "Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications", Environmental Modelling and Software, Vol. 15, Issue 1, pp. 101-124, 2000.
[42] Veenadhari, S., "2007, "Crop productivity mapping based on decision tree and Bayesian classification", M.Tech Thesis, submitted to Makhanlal Chaturvedi National University of Journalism and Communication, Bhopal.
[43] Patcharanuntawat, P., Bhaktikul, K., Navanugraha, C. and Kongjun, T., "Optimization for cash crop planning using genetic algorithm: A case study of upper Mun Basin, Nakhon Ratchasima province", 4th INWEPF Steering Meeting and Symposium, Paper 2-07: pp. 2-13, 2007.
[44] Camps-Valls, G., Gómez-Chova, L., Calpe-Maravilla, J., Soria-Olivas, E., Martín-Guerrero, J. D. and Moreno, J., "Support vector machines for crop classification using hyperspectral data", In: Pattern Recognition and Image Analysis, Springer, Berlin, Heidelberg, pp. 134-141, 2003.
[45] Tripathi, S., Srinivas, V.V. and Nanjundiah, R.S., "Downscaling of precipitation for climate change scenarios: a support vector machine approach", Journal of Hydrology, Vol. 330, Issue 3, pp. 621-640, 2006.
[46] Karimi, Y., Prasher, S.O., Patel, R.M. and Kim, S.H., "Application of support vector machine technology for weed and nitrogen stress detection in corn", Computer and Electronics in Agriculture, Vol. 51, pp. 99–109, 2006
[47] Goel, P.K., Prasher, S.O., Landry, J.A., Patel, R.M., Bonnell, R.B., Viau, A.A. and Miller, J.R., "Potential of airborne hyperspectral remote sensing to detect nitrogen deficiency and weed infestation in corn", Computur and Electronics in Agriculture, Vol. 38, Issue 2, pp. 99-124, 2003.
[48] Tellaeche, A., BurgosArtizzu, X.P., Pajares, G. and Ribeiro, A., "A vision-based hybrid classifier for weeds detection in precision agriculture through the Bayesian and Fuzzy k-Means paradigms", In: Innovations in Hybrid Intelligent Systems, Springer, Berlin, Heidelberg, pp. 72-79, 2007
[49] BCC Research, "Biopesticides: the global market", BCC Research Report, CHM029C, 2010.
[50] Abdullah, A., Brobst, S., Pervaiz, I.. Umar, M. and Nisar, A., "Learning dynamics of pesticide abuse through data mining", Australasian Workshop on Data Mining and Web Intelligence, Dunedin, New Zealand, 2004.
[51] Abdullah, A. and Hussain, A., "Data mining a new pilot agriculture extension data warehouse", Journal of Research and Practice in Information Technology, Vol. 38, Issue 3, pp. 229–249, 2006.
[52] Schatzki, T.F., Haff, R.P., Young, R., Can, I., Le, LC. and Toyofuku, N., "Defect detection in apples by means of X-ray imaging", Transactions of the American Society of Agricultural Engineers, Vol. 40, Issue 5, pp. 1407–1415, 1997.
[53] Shahin, M.A., Tollner, E.W. and McClendon, R.W., "Artificial intelligence classifiers for sorting apples based on watercore", Journal of Agricultural Engineering Research, Vol. 79, Issue 3, pp. 265–274, 2001.
[54] Urtubia, A., Perez-Correa, J.R., Meurens, M. and Agosin, E., "Monitoring large scale wine fermentations with infrared spectroscopy", Talanta, Vol. 64, Issue 3, pp. 778–784, 2004. doi:10.1016/j.talanta.2004.04.005.
[55] Mucherino, A. and Urtubia, A., "Consistent biclustering and applications to agriculture", IBAI Conference Proceedings, Proceedings of the Industrial Conference on Data Mining (ICDM10), Workshop Data Mining in Agriculture (DMA10), Springer, pp. 105–113, 2010.
[56] Urtubia, A., Pérez-Correa, J.R., Soto, A. and Pszczolkowski, P., "Using data mining techniques to predict industrial wine problem fermentations", Food Control, Vol. 18, Issue 12, pp. 1512-1517, 2007.
[57] Riul, A. Jr,, de Sousa, H.C., Malmegrim, R.R., dos Santos, D.S. Jr., Carvalho, A.C.P.L.F., Fonseca, F.J., Oliveira, Jr.O.N. and Mattoso, L.H.C., "Wine classification by taste sensors made from ultra-thin films and using neural networks", Sens Actuators, Vol. B98, pp. 77–82, 2004.
[58] Ahmadi, H., Golian, A., Mottaghitalab, M. and Nariman-Zadeh, N., "Prediction model for true metabolizable energy of feather meal and poultry offal meal using group method of data handling-type neural network", Poultry Science, Vol. 87, Issue 9, pp. 1909–1912, 2008. doi:10.3382/ps.2007-00507. ISSN 0032-5791.
[59] Ahmadi, H., Mottaghitalab, M., Nariman-Zadeh, N. and Golian, A., "Predicting performance of broiler chickens from dietary nutrients using group method of data handling-type neural networks", British Poultry Science, Vol. 49, Issue 3, pp. 315–320, 2008. doi:10.1080/00071660802136908.
[60] Chedad, A., Moshou, D., Aerts, J.M., Van Hirtum, A., Ramon, H. and Berckmans, D., "Recognition system for pig cough based on probabilistic neural networks", Journal of Agricultural Engineering Research, Vol. 79, Issue 4, pp. 449–457, 2001.
[61] Tahmoorespur, M. and Ahmadi, H., "Neural network model to describe weight gain of sheep from genes polymorphism, birth weight and birth type", Livestock science, ISSN 1871-1413, 2012.
[62] Ahmadi, H. and Golian, A., "The integration of broiler chicken threonine responses data into neural network models", Poultry Science, Vol. 89, Issue 11, pp. 2535–2541, 2010. doi:10.3382/ps.2010-00884. ISSN 0032-5791.
[63] Fernandez Pierna, J.A., Baeten, V., Michotte Renier, A., Cogdill, R.P. and Dardenne, P., " Combination of support vector machines (SVM) and near-infrared (NIR) imaging spectroscopy for the detection of meat and bone meal (MBM) in compound feeds", Journal of Chemomology Vol. 18, pp. 341–349, 2004.