Volume 5, Issue 5, May 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Different Techniques Used in Data Mining in Agriculture

Jyotshna Solanki, Prof. (Dr.) Yusuf Mulge
PDM College of Engineering & Technology
Bahadurgarh, Haryana, India

Abstract: Data mining is a process of extracting/equating the meaningful data from the large database by using different tools and techniques. Data mining (sometimes called information or knowledge discovery) is the process of evaluating data from different outlooks and summarizing it into useful information. This paper includes the different techniques and approaches like classification, association and prediction.

Keywords: Data Mining, Techniques, Applications etc.

I. INTRODUCTION

Data mining is a process of extracting/equating the meaningful data from the large database by using different tools and techniques. Data mining (sometimes called information or knowledge discovery) is the process of evaluating data from different outlooks and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of diagnostic tools for analyzing data. It allows users to investigate data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or forms among dozens of fields in large relational databases.Data mining or knowledge discovery in databases (KDD) is an interdisciplinary field where we integrate techniques from different fields including data base systems, statistics, mathematics, high performance computing, artificial intelligence and machine learning.

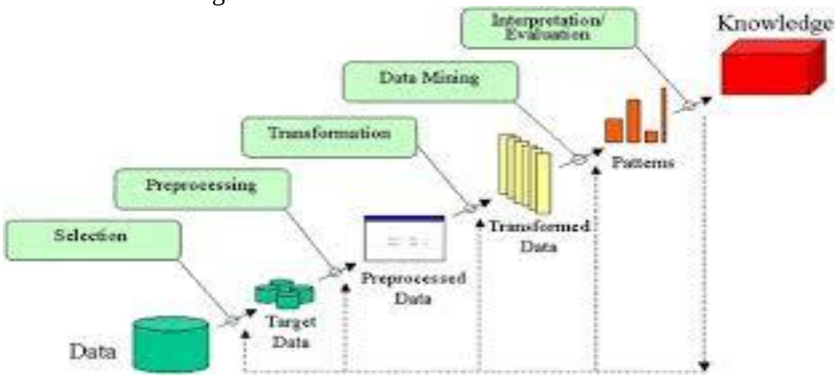

Fig 1 Basic idea about KDD

Data/Information is any facts, numbers, or text that can be processed by a computer. Organizations are accumulating vast and growing amounts of data in different formats and different databases. Information can be converted into knowledge about historical patterns and future trends. The patterns, associations, or relationships among all this data can provide information.
This includes:

- Operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- Nonoperational data, such as industry sales, forecast data, and macro-economic data
- Meta Data - data about the data itself, such as logical database design or data dictionary definitions.

II. DATA WAREHOUSES

Data Warehouses are the collection and managing the huge database at one palace so that in future it can be access easily. In the database management system the statistics should be self-determining, resourceful access, data reliability, security, and concurrent access and help to recover from huge disaster. Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized data management and retrieval the information/ knowledge. Data warehousing represents visualization of maintaining a central repository of all administrative data. Centralization of data is needed to maximize user access and analysis.

While far-reaching information technology has been evolving separate operation and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- *Classes*: Stored data is used to locate data in predetermined groups.
- *Clusters:* Similar Data items are grouped according to logical relationships.
- *Associations*: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- *Sequential patterns*: Data is mined to antedate performance arrangements and trends.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software .Present the data in a useful format, such as a graph or table.

III. DATA MINING IN AGRICULTURE

Data mining in agriculture is a very recent research topic. It consists in the application of data mining techniques to agriculture. This data mining techniques used in agriculture for prediction of problem, disease detection, optimizing the pesticide and so on .Recent technologies are nowadays able to provide a lot of information on agricultural-related activities, which can then be analyzed in order to find important information and to collect relevant information. This data mining techniques are used for disease detection, pattern recognition by using multiple application. Data mining is about to identify the similarities between searching the valuable business information from the large database systems such as finding linked products in gigabytes of store scanner data or the mining a mountain for a vein of valuable dataset. Both kind of processes required either shifting through an immense amount of material, or to perform the search intelligently so that exactly match will be performed. Data mining can be done on a database whose size and quality are sufficient. The technology of data mining can generate new business opportunities by providing these capabilities:

- *Automated prediction and analysis of various trends and behaviors* - Data mining itself automate the process by obtaining the predictive information from large databases. It first setup the questions and then provides the relative solutions. A typical example of such a predictive system is in the marketing field. Data mining uses data on the historical promotional mailings to capture the targets effectively so that the maximum return from market will be achieved. Other predictive problems include the detection of bankruptcy and or the frauds.
- *Another application of data mining is automated discovery of historical patterns dynamically*. The presented Data mining system is able to sweep over the databases to identify the hidden patterns. One of such example of pattern discovery is the analysis of retail sales data to identify the seemingly unrelated products so that the effective purchase can be done. Other pattern discovery analysis includes the detection of fraudulent credit card as well as the transactions to identify the anomalous data.

Data mining techniques are able to get the benefits of automation on existing software and hardware platforms that can be implemented on new systems which can be upgraded and new products can be developed. When data mining tools are defined on high performance parallel systems, they can be analyzed with massive databases in minutes. Faster processing is required in such system to derive the effective results from complex systems. High speed processing and accurate outcome from the system makes it possible for users to analyze large set of information. Larger databases, in turn, gives more improved predictions.

- Agricultural organizations store huge amounts of data in the form of crop databases. Trends in these databases can be identified using data mining practices, which sort and model the data in order to arrive at a conclusion. The data mining applications present the data in the form of data marts.
- In the agricultural industry, however, the lack of standard vocabulary has hindered the process of data mining to a certain extent. This could lead to unnecessary problems, during the process of data mining. The increase in the use of standardized terms will reduce the percentage of errors in the data mining process.

IV. TECHNIQUES UESD IN DATA MINING

Data mining algorithms primarily based on statistical techniques and machine learning algorithms. During the design of models, algorithms are used. Not only algorithms are important in this context but also data itself is critical there are two forms of data analysis and study so that we can examine the things and gets the knowledge that can be used for extract models describing important classes or predict future data trends. These data analysis help us to provide a better understanding of large data these two forms are as follows:

- *Classification*: In Classification prophesies unconditional and calculation models predicts continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.
- *Prediction*: In prediction which is used to predict the data according to the given data. Following are the examples of cases where the data analysis task is Prediction For example the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this example we are bother to predict a numeric value. Therefore the data analysis task is example of numeric prediction.

V. MAJOR ISSUSES IN CLASSIFICATION AND PREDICTION

The major issue is preparing the data for Classification and Prediction. Preparing the data involves the following activities:

- *Data Cleaning* - Data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.
- *Relevance Analysis* - Database may also have the irrelevant attributes. Correlation analysis is used to know whether any two given attributes are related.
- *Data Transformation and reduction* - The data can be transformed by any of the following methods.
- *Normalization* - The data is transformed using normalization. Normalization involves scaling all values for given attribute in order to make them fall within a small specified range. Normalization is used when in the learning step, the neural networks or the methods involving measurements are used.
- *Generalization* -The data can also be transformed by generalizing it to the higher concept. For this purpose we can use the concept hierarchies.

VI. RELEATED WORK

- In Year 2010, Li Zhanfeng performed a work," An Unattended Detection Method of Soil Respiration". This method uses proposed utility line peak extraction method to extract the CO2 concentration in the soil respiration chamber, and sets up a mathematical model of soil CO2 concentration. The system equipment is independently developed and it is verified that the extraction method is feasible with the site automatically detection [1].
- In Year 1994, Metin Akay performed a work, "Automated Noninvasive Detection of Coronary Artery Disease Using Wavelet-Based Neural Networks". This study examines the utility of neural networks for detecting coronary artery disease noninvasively by using clinical examination variables and extracting useful information from the diastolic heart sounds associated with coronary occlusions. It has been widely reported that coronary stenosis produce sounds due to the turbulent blood flow in these vessels [2].
- In Year 1991, Xavier SIMON performed a work, "Plot Analysis for Crop Identification and Disease Detection". Most methods and algorithms use satellite images as a main or exclusive source of information. Author describes a crop identification method which uses geographic information to identify agricultural plots on a time succession of images, and analyses its history through several years of cultivation. The pixels of each plot are considered as a statistical population and the distribution parameters of the distributions of their radiometric values are calculated for each wavelength channel. Author use these parameters to perform discriminant analyses bearing on the test-site plots and, at a later stage, on the surrounding plots. The results of this classification are the first step of a deductive method which allows a precise and early crop identification and further investigation of other agricultural parameters, such as damage caused by plant disease [3].
- I.V.Bragin performed a work, "Passive Microwave Complex of Remote Sensing for the Detection of Objects in the Soil". The paper deals with some grounding in theory of a passive radiometric complex designed to detect from a helicopter metal and non-metal objects in the soil at the depth of several meters [4].
- Takashi Kido performed a work," Haplotype Pattern mining & Classification for detecting disease associated Site". In this paper, Author introduces a new method for effective haplotype pattern mining to detect disease associated mutations. Using this procedure, Author can discover some of the new disease associated SNPs, which cannot be detected by traditional methods. Author will introduce a powerful tool for implementing this procedure with some worked examples [5].
- In Year 2004, Bernard J. Vigier performed a work," Narrowband Vegetation Indexes and Detection of Disease Damage in Soybeans". A portable narrowband spectroradiometer was used to detect sclerotinia stem rot infection, caused by the fungus Sclerotinia sclerotiorum in soybeans. Increasing levels of fungal inoculum were used to cause a gradient of disease infection in the field. A new field approach is suggested for the investigation of plant damage with narrowband spectroradiometry [6].

VII. DATA MINING APPROACHES

- *Decision Tree*: A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. In decision analysis a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of 3 types of nodes:

1. Decision nodes - commonly represented by squares
2. Chance nodes - represented by circles
3. End nodes - represented by triangles

- Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. If in practice decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities. Decision trees, influence diagrams, utility functions, and other decision analysis tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or management science

- *k-Nearest Neighbors algorithm*: In pattern recognition, the *k*-Nearest Neighbors algorithm (or *k*-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression:
- In *k-NN classification*, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor.
- In *k-NN regression*, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbors.
- *k-NN* is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learnin*g* algorithms. Both for classification and regression, it can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of $1/d$, where *d* is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for *k*-NN classification) or the object property value (for *k*-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A shortcoming of the *k*-NN algorithm is that it is sensitive to the local structure of the data. The algorithm has nothing to do with and has not to be confused with *k*-means, another popular machine learning techniques
- *Naive Bayes classifiers:* In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong naive independence assumptions between the features. It was introduced under a different name into the text retrievalcommunity and remains a popular baseline method for text categorization, the problem of judging documents as belonging to one category or the other with word frequencies as the features. With appropriate preprocessing, it is competitive in this domain with more advanced methods including support vector machines linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. In the statistics and computer science literature, Naive Bayes models are known under a variety of names, including simple Bayes and independence Bayes

VIII. APPLICATIONS

- *Prediction of problematic wine fermentations*: In this application the expectation will be computed and then solution is triggered. For say Wine is widely produced 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 we were able to predict how the fermentation is going to be at the early stages of the process, we could interfere with the process in order to guarantee a regular and smooth fermentation. Fermentations are nowadays studied by using different techniques,
- *Detection of diseases from sounds issued by animals*: The detection of animal's diseases in farms can impact positively the productivity of the farm, because sick animals can cause infections. Moreover, the early detection of the infections can allow the farmer to cure the animal as soon as the disease appears. Sounds issued by pigs can be analyzed for the detection of diseases.

IX. CONCLUSION

Data Mining is an integral component of all the databases for selecting the meaningful information from the from data. This paper summarizes the different techniques of data mining and also highlighting the various applications used in data mining. And also includes the various issues in data mining in agriculture. This Paper Deals with the techniques used in agriculture for exacting the meaningful data for relevant information. Data mining in agriculture is a very recent research topic. It consists in the application of data mining techniques to agriculture. This data mining techniques used in agriculture for prediction of problem, disease detection, optimizing the pesticide and so on

REFERENCES

[1] Li Zhanfeng," An Unattended Detection Method of Soil Respiration", 978-1-4244-7161-4/10©2010 IEEE

[2] Metin Akay," AUTOMATED NONINVASIVE DETECTION OF CORONARY ARTERY DISEASE USING WAVELET-BASED NEURAL NETWORKS", 0-7803-2050-6@1994 IEEE

[3] Xavier SIMON," PLOT ANALYSIS FOR CROP IDENTIFICATION AND DISEASE DETECTION", CH2971-0/91/0000-1927@1991 IEEE

[4] I.V.Bragin ," PASSIVE MICROWAVE COMPLEX OF REMOTE SENSING IFOR THE DETECTION OF OBJECTS IN THE SOIL".

[5] Takashi Kido," Haplotype Pattern Mining & Classification for detecting disease associated Site".

[6] Bernard J. Vigier," Narrowband Vegetation Indexes and Detection of Disease Damage in Soybeans", IEEE GEOSCIENCE AND REMOTE SENSING LETTERS 1545-598X/04© 2004 IEEE

[7] Anshuk.a Srivastava," Development of a Sensor for Automatic Detection of Downey Mildew Disease", Proceedings of the 2004 lntenatlonal Conference on Intelligent Mechatronics and Automation 0-7803-8748-1/04/@2004 IEEE

[8] Ting Li," Prediction and Early Warning Method for Flea Beetle Based on Semi-supervised Learning Algorithm", Fourth International Conference on Natural Computation 978-0-7695-3304-9/08 © 2008 IEEE

[9] Zhang Lei," Characteristic of lead content in soil and crop in Zhaoyuan gold mine area", 2012 International Conference on Biomedical Engineering and Biotechnology 978-0-7695-4706-0/12 © 2012 IEEE

[10] A. K. Tripathy," Data Mining and Wireless Sensor Network for Groundnut Pest/Disease Precision Protection".