An Approach towards Crop Management using DM and Image Processing

Suraj Balwan
M.Tech, BE(CSE)
Modern Education Society's
College of Engineering
Pune-411001,India

Ajinkya Kunjir
BE(COMPUTER SCIENCE)
Modern Education Society's
College of Engineering
Pune-411001,India

Mahesh Malwade
M.Tech, BE(CSE)
Modern Education Society's
College of Engineering
Pune-411001,India

ABSTRACT

The recent development and improvements in fields of computer science and technology has led to several innovations and inventions with the help of trending Algorithms, Domains, Techniques and tools. Data science is the domain of computer science comprising of Big Data and Data Mining, where Data Mining is the backbone of data science and has provided a base to Big Data which is now a trend in data science. Applications of Data Mining range from fields of Healthcare, Military, Aviation, Business systems and Agriculture. Agriculture, an application and area of data mining is a very recent research topics. Trending methods and techniques are nowadays able to produce and supply a large amount of data and knowledge on farming and agriculture activities, which can further be analysed in order to uncover some important matter out of it. The paper explains the proposed system which uses the combined technology of Data mining and Image processing. The data mining algorithms such as Random forest, Decision tree induction's J48 and others are compared on a testing platform on basis of accuracy and other performance factors. Algorithm which outclasses the comparative study is selected for direct implementation in the system and its database. Image processing is used for identifying the crops, and process their image to forward it to further analysis. The crops will be classified by the Data Mining algorithms on basis of labels such as weak, strong, or mid-healthy. Data Mining and Image processing, technologies when combined together in one system can exploit the levels of parallelism and can achieve an enhanced faster rate of classification and processing. The records and tabular structure of the plants or agriculture data can be visualized using Data Mining Visualization techniques such as 2D/3D Graphs, Pie charts, tree graph etc.

General Terms

Data Mining and techniques

Keywords

Image Processing, Image Recognition, Data Mining, Sensors, Data Warehouse, J48, Random Forest, Visualization.

1. INTRODUCTION

Computer Science is a field consisting and made up of many domains such as Artificial Intelligence, Software design and methodologies, Internet of Things (IoT), Fuzzy logics, etc. Data Mining and Big Data are two new areas which are trending in the domain of data science. Data mining is an important domain for many reasons. KDD, which is "Knowledge Discovery from Data" is the basic and backbone process in this field of identifying and extracting data from the huge databases. Data Mining is a sub-step of this process. According to definitions given in the articles, "Data Mining is a process in which the selected dataset or database is analysed and interesting patterns are evaluated out from that data. In KDD, new patterns of knowledge are identified, determined, and extracted from the data. Ankita Dewan, Meghna Sharma in their paper of 2015, stated that the areas like healthcare, banking and finance, business and IT companies have adopted Data mining as their provider technology. A lot of data is generated from this areas which is not sufficiently used for pattern prediction purpose [1]. The data which is not used or is wasted can be made useful by converting it into knowledge and can also be modelled using different Data Mining techniques and methods. Information or data can be defined as a collection of facts or records which were captured from events. Companies are generating and storing huge amounts of data in respective specific formats and various databases. Information when organized well can be converted into knowledge. Knowledge comprises of different views, ideas and experiences of the people. Interesting information and associative relations can be found out from these patterns. This includes Operational data such as customer record, sales, marketing, finance, etc. Also included are informational data, such as company data, weather forecast, product data and Meta Data (data that describes the data in datasets).

1.1 Data Warehouse

Data warehouse can be defined as the centralized storage for all the data which is extracted from external data sources from the data source layer. Jyotshna Solanki et.al in their paper explained the primary purpose of data warehouse is to generate efficient reports and analysis to support decision making process [2]. Data stored in data warehouse is in a three dimensional (multidimensional) form. Basically the data is stored in form of 3D cubes and cuboid lattices. The main characteristics of data warehouse are as follows:

1. Subject-Oriented: The data is stored in a department wise fashion in all the data-marts.
2. Integrated: The data to be extracted is first converted into the data warehouse format and is then stored in the data warehouse.
3. Time-Variant: The data can be entered in data warehouse by time intervals such as by days, months, years, etc.
4. Non-Volatile: The data in the data warehouse cannot be changed or modified. Old data can be discarded.
5. Multidimensional Storage: The data stored in data warehouse is in form of 3D cubes, lattices, etc.

Besides all these characteristics, Both Granular and summarized data can be stored in the data warehouse for enhanced analysis. There are two types of data warehouses, such as centralized and distributed. Data warehouse can provide a centralized storage for storing all the subject oriented data of agriculture and plants such as crop data, crop types, crop details, land details, etc. Data warehouse is well known for storing all the current as well as historic data.

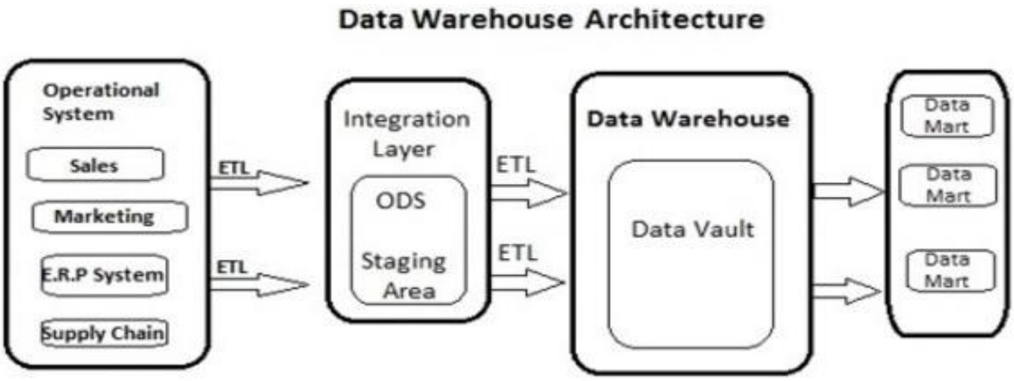

Fig 1: Data Warehouse Architecture

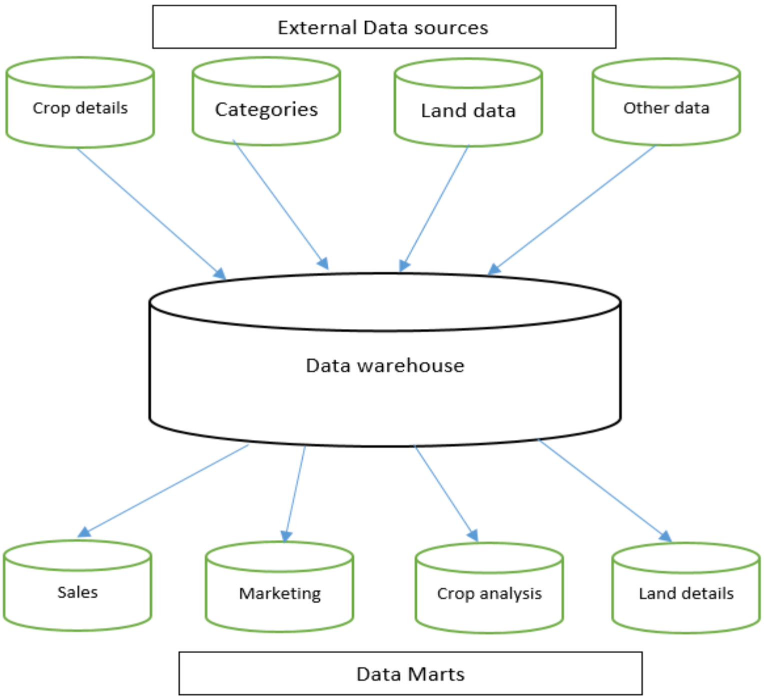

Fig 2: Data warehouse Process for Agriculture

From Fig 2, we can understand the layered architecture of the Data Warehouse. The layers consists of steps which describes the whole process starting with data extraction, data loading, data staging, and also the last stage is data presentation in form of reports, Querying and statistical analysis. The external data sources which are extracted from agricultural organizations and stored in a subject oriented fashion in the data warehouse. The data marts arrange the data extracted from the data sources.

2. LITERATURE SURVEY

2.1 Related Work in Agriculture using DM

The techniques of Data Mining are used in applications like MBA(Market Basket Analysis), Banking, Telecommunication and others.. Data mining also includes predictive and descriptive mining methods such as prediction, classification, regression and association rules. Data mining is used to construct a decision support system, Recommendation system, Multi-purpose system, and an expert system in fields of healthcare, military, forest management and agriculture. L Jagielska et.al in their paper stated the use of k-means methodology of data mining for classifying soil in combination with GPS [3]. Tellaeche et.al in their paper based on hybrid classifier model for weed detection in agriculture used k-means algorithm and data mining strategy for wine fermentation issue [4]. K. Verheyen et.al in their article stated the working and usage of fuzzy sets for yield predictions in agriculture. The fuzzy logics and neural networks are too complex to implement on massive datasets [5]. S.Veenadhari et.al in their paper described the use of data mining techniques such as decision tree analysis for prediction of crop productivity. Application which was stated for this prediction analysis was the influence of climatic factors on kharif and rabi crops production [6].

2.2 Data Mining Techniques

Data mining methods and its algorithms are based on principles of statistical and advanced techniques and is based on foundations of machine learning. Algorithms of Data Mining are used for performing analysis and pattern extraction from the selected data. Ajinkya Kunjir et.al in their survey paper in 2016 described the use of data analysis its approach to help us to provide a better exploration and understanding of large data, which are classifications and predictions [7]. Various predictive and descriptive data mining techniques are described below as follows:

1. Clustering: The process of grouping elements which possess a similar characteristics and features in a same group is called as clustering. The data scattered in the hyper-plane is selected by implementing various clustering algorithms on them and the similar data elements are grouped on the basis of properties. The various clustering algorithms are k-means, k-mediods, agglomerative clustering and hierarchical clustering.
2. Association Rules: The mapping of relationship between the items of a same transaction or a same database. The items are called as “Associations” because they share an interesting quality relations.
3. Classification: The process of identifying the class labels of the unknown tuple of data is called as classification of the data. The previously unknown data can be classified to their categories according to their class labels. The several classifier models Decision Tree induction, Naïve Bayes classifier, etc.
4. Prediction: The techniques which is used as a term in classification. The classifiers predict the value of an unknown variable which is also a dependent variable. The several prediction techniques include regression and classifier models.

3. PROPOSED SYSTEM

We propose an efficient multiclass system which can act as a decision support system for agricultural officers and farming officers for taking crucial conservative decisions. The assumed theory and motive behind this project is to capture digital images of the crops and plants using technologies such as image processing and image recognition. The Camera equipped with image recognition techniques and algorithms are kept focused on the plants and capture the reading and encapsulate them in a tabular format in the database. The data mining algorithms selected for comparative study are J48 and Random Forest, the best out of the two will be selected for implementation purpose on the tabular datasets to identify the class label of an unknown data tuple related to plants and crop data.

4. COMPARATIVE STUDY OF ALGORITHMS

WEKA which is also known as “Waikato Environment for Knowledge Analysis” is a packaged suite for machine learning algorithms. This framework is written in Java language and is licensed under general public license [8]. The framework is a plethora of several visualization tools and algorithms which are used for data analysis and operating on the datasets which are in .arff format. The agricultural datasets deployed by their organizations. The algorithms J48 and Random Forest were tested on agricultural datasets of the farming organizations. The algorithms were compared and tested on basis of accuracy and latency performance analysis. According to the test results “J48/Random Forest” Outperformed the performance analysis by giving out the best prediction accuracy.

4.1 J48

The successor of C4.5 is J48, which can also be said as the simple derivation of C4.5. It forms a binary tree. The decision tree approach is most useful and beneficial in issue of classification. Tree construction can be made successful using the classifier models or classification process. After finishing with the construction, each tuple in the database can be tested and classified.

Algorithm for Decision tree’s J48:

```
INPUT:
A                       //where, A is the training data

OUTCOME:
O                       // where, O is the decision tree

ATCREATE (*A)
{
O= NULL;                          // initialised as null
O= Create root node and label;
O= Addition of arc to each split node
For every arc added do
D= Apply spilt predicate to D and create rood node;
When reached the final stop point do
O’= Creation of leaf node with correct class label;
Else
O’= DTCREATE (D);
O= addition of O’ to arc;
}
```

J48 ignores the missing values while constructing a decision tree. Each item can be processed for calculating its value and prediction based on value of other records. The ideal strategy is to decompose the data into range based on the values of the attributes for the item found in training instances. Decision tree induction can be used to achieve classification for J48.

4.2 Random Forest

Random forest algorithm produces only two important variables which are variable importance and proximity measures. Variable importance is a difficult concept due to its communication among each other. The importance of variable can be calculated using Random Forest by observing the increase in prediction accuracy when the objects are permuted and others have been unchanged.

Steps of Random Forest Algorithm:

1. From the original data Draw N_{tree} samples
2. Grow an unpruned tree for each examples, make the alterations at each node besides selecting the best split among the predictors
3. Sample M_{try} instances randomly among the predictors and select the best split from those variables.
4. Classify the new data by combining the predictions of N_{trees} Samples.
5. Obtain the measure of error rate of the training data

The Proximity measures and matrix can be used to locate the position of elements i.e in form of (i, j), where both the parameters can be used to locate the elements. The algorithm described above is used for both regression and classification.

5. STUDY OF SENSORS

5.1 Humidity and temperature sensors

Crops or plants shade off peal and dry because of less amount of sunlight, water and pesticides. The main objective of this proposed system is to detect if a plant is healthy or not and whether it requires a proportional amount of utilities and resources to become healthy. The sensors used in this system are programmed to detect and sense the condition and parameters of the crop. The sensor technology is linked with the image processing and the data is converted into tabular form for data mining algorithms to act on it.

5.1.1 SEN13322 ROHS

The SparkFun Soil Moisture Sensor is an easy sensor technology for calculating the soil's moisture. plants. The sensor described in this paragraph is quite easy to use. The two big pads at the sensors mouth function as sensor's probes. The conductivity between the pads is more if the water in the soil is more. If these two things succeed then it results in low resistance and high signal. The most common occurring problem with soil moisture sensors is their less durability when they are exposed to the moisture.

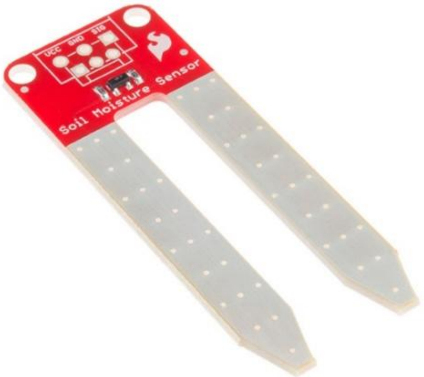

Fig 3: SparkFun Soil Moisture Sensor

6. WHEN DM MEETS IMAGE PROCESSING

Data mining is a technology in which potential and valid patterns are identified and extracted from the selected or target dataset. The patterns carved out of data are used for analysis and future prediction purpose. Image processing is a technology in which the images which are identified and extracted are processed by using different image processing algorithms to synthesize and smoothen the image. When data mining meets image processing, different technologies such as image mining, image retrieval, pattern recognition, image recognition are evolved. The patterns which are generated out from the data by the data mining algorithms are given further for processing to the IP system. Image mining is a technique in which patterns and information is identified and extracted from large volumes of data and converted into knowledge.

7. APPLICATIONS OF DM IN AGRICULTURE

7.1 Mushroom Grading

Prof. Navneet Goyal of BITS Pilani elaborated the mushroom grading and the related application in his document [9]. Data mining techniques and algorithms can be used for development of a classifier model for quality grading of mushrooms for achieving an accuracy similar to the accuracy achieved by inspection done by agriculture investigators.

7.2 Apple Pest Management

Apple proliferation is one of the important phytoplasma disease which infects the growth of apple. The disease symptoms in form of size reduction, silver leaf fungus, abnormal shape, discolouration, witches broom and fungal growth. Data mining and image processing has developed an apple proliferation system which can detect the symptoms and suggest preventive measures which can help reduce the fungal infection of plants and enhance their growth.

7.3 Pesticide Abuse

Researchers in their previous work stated that the maximization of crop yielding through pesticide policies can result in harmful and high usage of pesticides. The negative co-relation between crop yield and pesticide usage was detected. Excess in use of pesticides for plants can cause harmful impacts to crops. To overcome this solution Mr. Dilip Roy Chowdhury et.al in their paper described the working of Abdullah and Brobst in 2003, where they invented a clustering technique of data mining, in which clustering of data was done through RNR (Recursive Noise Removal) technique [10]. Clustering technique of Data Mining can be used to group together the pesticide elements to find out the total pesticide usage.

8. CONCLUSION

The paper discusses about all the data mining techniques and methods which can be used to identify and suggest measure to avoid disruption of crops and plants. The sections described above also tells us about the converging of image processing and data mining. The merging of these two technologies can provide efficient suggestions and technological help for building a crop management systems using data mining algorithms and image processing techniques. The paper also enlightens us about the humidity and temperature sensors which are used in this system for detecting the health and conditions of the crops. The algorithms such as Random Forest and J48 were selected for performance analysis and comparative study on agricultural datasets. The algorithm which outclasses the test will be selected for further implementation. The later sections describes the applications of data mining and previous works done by researchers and data mining practitioners in fields of agriculture.

9. REFERENCES

[1] Ankita Dewan, Meghna Sharma , "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification" , 2015 2ndInternationalConference on Computing for Sustainable Global Development (INDIACom), IEEE 2015.

[2] Jyotshna Solanki, Prof. (Dr.) Yusuf Mulge, "Different Techniques used in Data Mining",2015 International Journal of Advanced Research in Computer Science and Software Engineering.

[3] I. Jagielska, C. Mattehews, T. Whitfort, " A study in experimental evaluation of neural network and genetic algorithm techniques for knowledge acquisition in fuzzy classification systems, IEEE 2002.

[4] Tellaeche, A., BurgosArtizzu, X. P., Pajares, G., & Ribeiro, A. (2007), “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 (pp. 72-79). Springer Berlin Heidelberg

[5] K. Verheyen, D. Adriaens, M. Hermy, and S. Deckers, “High resolution continuous soil classification using morphological soil profile descriptions”, Geoderma, vol. 101, pp. 31-48, 2001.

[6] Veenadhari, S. 2007, “Crop productivity mapping based on decision tree and Bayesian classification”. Unpublished M.Tech Thesis submitted to Makhanlal Chaturvedi National University of Journalism and Communication, Bhopal.

[7] Ajinkya Kunjir, Harshal Sawant, Nuzhat F. Shaikh, “A Survey on Prediction of Multiple Diseases using Data Mining and Visualization Techniques”, IJCA December 2016.

[8] Wikispaces https//weka.wikispaces.com

[9] Prof. Navneet Goyal, “Data Mining Applications in Agriculture”, BITS PILANI.

[10] Dr. Dilip Roy Chowdhury, Subhashis Ojha, “An Emprical Study on Mushroom Disease Diagnosis : A Data Mining approach”, IRJET 2017.