Asia-pacific Journal of Convergent Research Interchange
Vol.2, No.4, December 31 (2016), pp. 25-33
http://dx.doi.org/10.21742/apjcri.2016.12.04

Implementation of Decision Based Fruits Protection System Using Classification and Clustering Techniques

Dong Jo Kim[1)], Deepa Sharma[2)]

Abstract

Automatic methods for the detection of plant diseases before hand is very important for absolute fruit protection. By using different data mining techniques we identify the diseases and take preventive measures to protect the fruit. Data mining techniques are the analytical tools that are used to extract the meaningful information from large data bases. This term paper highlights the use of data mining techniques like clustering and classifications to improve protection of fruit from diseases which includes certain parameters like weather conditions like temperature, amount of rainfall, humidity of air and leaf wetness classification technique helps us to identify to which set of category particular observations belong. Clustering technique is used to combine the results occurred.

Keywords: data mining, classification, clustering, disease prediction, fruit protection.

1. Introduction

Environmental changes and an unnatural weather change produce[1-3] challenges for farming generation. On the first side individuals[1-3]. have less unadulterated arable area, and on other expanding number of maladies and pests cause the utilization of considerably more chemicals[4]. Those chemicals connected in substantial amounts lead to soil pollution, and could jeopardize human wellbeing[5]. This issue can be dodged by utilizing chemicals at the opportune time, when minimal measure of them can smoother the solid pathogen or pest[6-9]. With a specific end goal to foresee the privilege time for concealment of pathogens numerous parameters must be prepared. For example, pathogen could infect fruit product species simply under specific conditions. Those ecological details could be particular climate conditions like temperature, measure of precipitation, humidity of air and leaf wetness. Other group of conditions is the presence of dynamic spores of a given pathogen. Farmers are trying to protect the fruits through their experience.

Received(September 19, 2016), Review Result(1st: October 7, 2016, 2nd: November 4, 2016), Accepted(December 10, 2016)

[1](Corresponding Author) Seoul Media Institute of Technology, 7F, 402 Worldcupbuk-ro, Mapo-gu, Seoul, Korea
email: sojudj@gmail.com
2) Department of Computer Science and Engineering, KL University
email: er.deepa.in@gmail.com

ISSN: 2508-9080 APJCRI

2. Related work

In some cases experience is insufficient. For exact expectation of conceivable organic product disease, and right time for organic product insurance data mining strategies can be utilized. Data mining is most valuable in an exploratory examination situation in which there are no faults and will give good results.It is a helpful exertion of people and PCs. Best results are accomplished by adjusting the learning of human specialists with the look abilities of PCs . Datamining comprise of two essential objectives, forecast(prediction) and description. Here we prefer for forecast. Expectation includes utilizing variables or fields in the information set to foresee obscure or future estimations of different variables of hobby. Datamining is rising exploration field in Horticulture plant security as well. For exceptionally late uses of data Mining procedures in agribusiness field distinctive. Datamining procedures are in use, for example, K-means, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN) and Suport Vector Machines (SVM). In our exploration and framework execution we will utilize new philosophies to anticipate conceivable contamination on the natural product plants. Consider that information are accessible from some point back to the past, where the comparing pathogen contamination has been recorded[2]. In each datamining strategy the preparation information gathered from some point back to the past is utilized as a part of terms of preparing which must be misused to figure out how to arrange the conceivable rise of sicknesses. This paper gives review of datamining strategies that can be utilized for expectation in a wide range of fields, and here creators propose model for expectation of conceivable natural product contamination.Farmers require an important amount of significant knowledge mined from its past and current data sets using special methods and processes. In order to extract the data we use different data mining techniques. “Data mining” is a process retrieving data that is stored in the database. When we apply the data mining techniques on the dataset, there should be a methodology that starts from the problem definition, then preprocessing, then we come to the datamining methods like[2] classification, clustering, association, trend analysis. Finally, the knowledge representation process the data available in the large database. One of the major issue farmers are facing is major growth in diseases. It is difficult to manage such huge amounts of data hence various tools and methods are used to manage lots of data. So data mining is applied. Data mining tools predict future trends and behaviour patterns, allowing farmers to make quick prevention, knowledge-driven and appropriate decisions.

Datamining instruments can answer numerous inquiries that generally were excessively

tedious in past, making it impossible to determine. They can utilize order method, for enquiring the method of a natural product characterstics, or utilize estimation and expectation procedure to foresee the probability of an assortment of results, for example, tainted and non contaminated. Apparatuses required for gathering information are MYSQL DATABASE, WEKA information mining instrument and so on. A decent bunching calculation preferably ought to deliver assembles with unmistakable non-covering limits, in spite of the fact that an immaculate partition can't regularly be accomplished by and by.

The Classification Models are mainly composed of C4.5 and ID3 algorithms these classification models are known as Decision Trees, from data. There are training set of records. The each and every record has a structure similarly it consists of any number of attribute pairs or value pairs. The attributes represent the category to which record it belongs. The issue is to decide a choice tree that on the premise of answers to addresses about the non-classification characteristics predicts effectively the estimation of the class property. Generally the classification trait takes just the qualities {true, false}, or achievement, failure}, or something equal. Regardless, one of its qualities will mean disappointment. They need those outcomes of estimations taken Toward masters for a couple gadgets. For every gadget we recognize what the quality for every estimation is and what was chosen, if to pass, scrap, or repair it. That is, we have a record with as non-unmitigated characteristics the estimations, as absolute trait the aura for the widget. The main document portrays the structure of the records. The second record gives the Training Set, and the third the Test Set. The basic ideas behind ID3 are that case. In the decision tree each node corresponds to a non-categorical attribute and every circular segment to a conceivable estimation of that characteristic. A leaf of the tree determines the normal estimation of the downright trait for the records portrayed by the way from the root to that leaf. This defines what a Decision Tree is. In the choice tree at every hub ought to be related the non-straight out quality which is most enlightening among the traits not yet considered in the way from the root. This builds up what a "Decent" choice tree is. Entropy is utilized to quantify how enlightening is a hub. This characterizes what we mean by "Great". Coincidentally, this idea was presented in Information Theory. C4.5 is an augmentation of ID3 that records for distracted qualities, consistent trait esteem ranges, pruning of choice trees, guideline determination, et cetera. We can utilize this idea of addition to rank credits and to fabricate choice trees where at every hub is found the property with most noteworthy increase among the properties not yet considered in the way from the root. The plan of this requesting are twofold To make little choice trees with the goal that records can be distinguished after just a couple questions. To coordinate a sought after negligibility of the procedure spoke to by

the records being viewed as C4.5 presents various expansions of the first ID3 calculation. In building a choice tree we can manage preparing sets that have records with obscure trait values by assessing the addition, or the increase proportion, for a property by considering just the records where that characteristic is characterized. In utilizing a choice tree, we can organize records that have darken attribute values by evaluating those probability of the Different time allows impacts. The consider about systems to assess the slip execution of a Decision tree will be presumably a really propelled to practically undergrad courses. An mix at order level will be finished the middle of these classifiers to get those best multi-classifier methodology Furthermore exactness for every information set. Machine learning is firmly identified with and regularly covers with computational measurements; an order which likewise centres in expectation making using PCs. It has solid binds to scientific improvement, which conveys strategies, hypothesis and application spaces to the field. Machine learning is used in an extent of figuring endeavours where sketching out and programming express estimations is infeasible. Illustration applications incorporate spam sifting, optical character acknowledgment (OCR), web indexes and PC vision. Machine learning is at times conflated with information mining, where the last sub-field concentrates more on exploratory information examination and is known as unsupervised learning. ID3 is the antecedent to the C4.5 calculation, and is regularly utilized as a part of the machine learning and characteristic dialect handling areas.Figure the entropy of each quality utilizing the information set S. Part the set S into subsets utilizing the characteristic for which entropy is least (or, identically, data addition is maximum).Make a choice tree hub containing that quality. Recurse on subsets utilizing remaining properties. Choice tree components are the Drawn from left to right, a choice tree has just blasted hubs (part ways) however no sink hubs (merging ways). Thus, utilized physically, they can become enormous and are then frequently difficult to draw completely by hand. Generally, choice trees have been made physically - as the aside illustration indicates - albeit progressively, specific programming is utilized. The choice tree can be linearized into choice standards, where the result is the substance of the leaf hub, and the conditions along the way shape a conjunction in the if statement. All in all, the tenets have the structure if condition1 and the condition2 and the condition3 then the result. Choice tenets can likewise be produced by developing affiliation rules with the objective variable on the privilege.

3. Building the Classifier or Model

This stride is the learning step or the learning stage.

In this stride the order calculations assemble the classifier. The classifier is worked from the preparation set made up of database tuples and their related class names. Each tuple that constitutes the preparation set is alluded to as a classification or class. These tuples can likewise be alluded to as test, article or information focuses.

4. Classification and Prediction Issues

The real issue is setting up the information for Arrangement and Forecast. Setting up the information includes the accompanying exercises:

Data Cleaning – Information cleaning includes evacuating the commotion and treatment of missing qualities. The clamor is evacuated by applying smoothing procedures and the issue of missing qualities is explained by supplanting a missing worth with most regularly happening esteem for that characteristic.

Relevance analysis – Database may likewise have the superfluous characteristics. Connection examination is utilized to know whether any two given properties are connected.

Data Transformation and reduction – The information can be changed by any of the accompanying strategies.

Normalization – The information is changed utilizing standardization. Standardization includes scaling all qualities for given ascribe so as to make them fall inside of a little indicated range. Standardization is utilized when as a part of the learning step, the neural systems or the strategies including estimations are utilized.

Generalization – The information can likewise be changed by summing it up to the higher idea. For this reason we can utilize the idea chains of command.

5. Literature Survey

Jin HaiYue , Song Kai proposed in their research work the main aim is to identify disease in crops using IBLE algorithm.This algorithm consists of four parts i.e., the initial set; construction rules algorithm; build decision tree algorithm; type determination algorithm. In the

intial settings example set of features are considered.Certain rules are formed to construct an algorithm and decision trees like ID3 etc are used and finally the required output is reached.

6. Existing system

In this system the data about weather and other climatic conditions are stored in the database through automatically or through manually.The data whenever required is retrieved from the database. For the detection of disease can be done in four stages[6].

1.Data collecting
2.Data pre processing
3.Data processing
4.prediction

Data Collecting: For the prediction of possible fruit and plant disease infection weather data and data about active spores are important. Here they proposed[6] a model in such a way that two key devices are the device which collect the data are automatic weather stations and spore traps.The data collected in this is monitored on site or transferred to server. Automatic stations will measure and send data on every fifteen minutes. For manual stations measurements will be carried out on eight hours. Spore traps will be examined three times a day and it is a electronic microscopic observation. Parameters from both sorts of stations will be spared in framework database. From programmed stations report will be send through the system, consequently. Information from manual station must be entered by human. Information accumulation can be robotized in the event that we have more programmed stations with spore traps on them. Information from the electronic magnifying instrument about recognized spores will be spared in the same database as meteorological information. More spore traps give better scope. One spore trap will be placed close to the particular plant. In that way examination can be altogether quicker, in light of the fact that toward the begin we kill spores that are not particular for ailment assaulting that plant. Spore trap will be analyzed by the phytopathology. At this stage, the information accumulation stage is over. By this we mean on the information got in the present time. For successful expectation, information from the past ten and more years will be entered in the database. Data about distinguished contaminations on the field was acquired from agriculturists and from authority bodies accountable for checking the event of illness.

Data Preprocessing: Information preprocessing must be connected on both meteorological and information from electronic microscope. Information preprocessing is regularly dismissed however vital step in the datamining process. Information gathering strategies are frequently lightly controlled, bringing about out-of-range values, impossible data combinations missing qualities, and so on.

Data Processing: In this stage the last information set sorted out in preprocessing stage will be prepared by various datamining systems. Information about recognized spores will be characterized by collecting of sicknesses that can cause. Close to this, spores for every pathogen are ordered in two sorts. First type is dynamic spores, which means that such spores could bring about ailments in proper climate conditions. Second one is detached spores. Detached spores can't bring about contamination, although climate conditions are satisfied. Information gathered from the meteorological stations are progressively various.

Prediction: For disease infections all said parameters must have values in particular reach. In light of grouped information from the preparation model, we can anticipate if proper conditions for conceivable disease are fulfilled. From this minute we utilize made prepared model for forecast. New set of examples gave from the meteorological stations furthermore, research center will be utilized for test dataset creation. Test dataset has the same structure like preparing dataset. For the class values we can include two sorts of qualities. In the first place is question mark that shows that we don't know which class quality is suitable for that arrangement of information. In the second case we can foresee the class esteem naturally, and information our forecast. In the wake of entering all the present estimations of the cases in the dataset, forecast can begin. In forecast stage we utilize our spared preparing model.

Methodology: Gather the fruit details like temperature, rainfall, humidity, precipitation, disease, color, seed treatment etc. Apply data mining techniques like clustering, classification, association rules and segregate them into various categories. Based on results predict[6] the issues that lead to their poor growth thereby farmers can take effective measures. In order to implement this initially the details related fruits are stored in the MYSQL database. Using WEKA software clustering is done and based on the result analysis and protection of the fruit is implemented.

7. Algorithm

k-means is one of the least difficult unsupervised learning algorithms that take care of the well known clustering problem. The system takes after a basic and simple approach to arrange a given information set through a specific number of bunches (accept k groups) settled apriori. The primary thought is to characterize k focuses, one for every bunch. These focuses ought to be put cunningly as a result of various area causes diverse result. Along these lines, the better decision is to place them however much as could reasonably be expected far from each other. The following step is to take every point having a place with a given information set and partner it to the closest focus. At the point when no point is pending, the initial step is finished and an early gathering age is finished. As of right now we have to re-ascertain k new centroids as barycenter of the groups coming about because of the past step. After we have these k new centroids, another tying must be done between the same information set focuses and the closest new focus. A circle has been created. As an aftereffect of this circle we might see that the k focuses change their area regulated until no more changes are done or as such focuses don't move any more. At long last, this calculation goes for minimizing a target capacity know as squared mistake capacity given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (||x_i - v_{i||})^2$$

where

||xi-vj|| is the Euclidean distance between xi and vj

'Ci' is the number of data points in ith cluster

'c' is the number of cluster

8. Conclusion

Agricultural production is a very vast and toughest job. One of the most unpredictable and complex task is crop protection. The key factor for successful chemical fruit protection from diseases and pests is nothing but the right time. This means that the selection of chemicals is not as complex as timing determination for protection. Early fruit disease detection has a lot of benefits. From the angle of farmers, methods like suggested one provide important information

for successful chemical protection. Second benefit for the farmers is economical. They can save money if they reduce numbers of chemical treatments. This is because model indicates when conditions for diseases development are not fulfilled. In that case chemical treatment is not needed. From the perspective of healthy foods, reduced number of chemical treatments is very important. With appropriate detection and prediction we could get successful chemical protection and healthy food.

References

[1] S. Freilich, S. Lev, I. Gonda, E. Reuveni, V. Portnoy, E. Oren, M. Lohse, N. Galpaz, E. Bar, G. Tzuri, G. Wissotsky, A. Meir, J. Burger, Y. Tadmor, A. Schaffer, Z. Fei, J. Giovannoni, E. Lewinsohn, and N. Katzir, Systems approach for exploring the intricate associations between sweetness, color and aroma in melon fruits, BMC Plant Biology, **(2015)**, Vol.15, No.71.

[2] M. Khanmohammadia, F. Karamia, A. Mir-Marquésb, A. Bagheri Garmarudia, S. Garriguesb, and M. Guardiab, Classification of persimmon fruit origin by near infrared spectrometry and least squares-support vector machines, Journal of Food Engineering, **(2014)**, Vol.142, pp.17-22.

[3] J. A. S. Almeida, L. M. S. Barbos, A. A. C. C. Pais, and S. J. Formosinho, Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering, Chemometrics and Intelligent Laboratory Systems, **(2007)**, Vol.87, pp.208-217.

[4] G. Fan, J. Zha, R. Du, and L. Gao, Determination of soluble solids and firmness of apples by Vis/NIR transmittance, Journal of Food Engineering, **(2009)**, Vol.93, pp.416-420.

[5] R. Feugel, R. Carle, and A. Schieber, A novel approach to quality and authenticity control of fruit products using fractionation and characterisation of cell wall polysaccharides, Food Chem., **(2004)**, Vol.87, No.1, pp.141-150.

[6] A. M. Gómez-Caravaca, R. M. Maggio, V. Verardo, A. Cichelli, and L. Cerretani, Fourier transform infrared spectroscopy – Partial Least Squares (FTIR – PLS) coupled procedure application for the evaluation of fly attack on olive oil quality, Lebensmittel-Wissenschaft & Technologie, **(2013)**, Vol.50, No.1, pp.153-159.

[7] J. Lutsa, F. Ojeda, R. Van de Plas, B. De Moor, S. Van Huffel, and J. A. K. Suykens, A tutorial on support vector machine-based methods for classification problems in chemometrics, Anal. Chim. Acta, **(2010)**, Vol.665, No.2, pp.129-145.

[8] B. M. Nicolaï, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K. I. Theron, and J. Lammertyn, Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: a review, Postharvest Biol. Technol., **(2007)**, Vol.46, No.2, pp.99-118.

[9] J. S. Ribeiro, T. J. Salva, and M. Ferreira, Chemometric studies for quality control of processed Brazilian coffee using DRIFTS, J. Food Quality, **(2010)**, Vol.33, No.2, pp.212-227.