Computers and Electronics in Agriculture 150 (2018) 418–425

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

# Early cherry fruit pathogen disease detection based on data mining prediction

Milos Ilic, Sinisa Ilic, Srdjan Jovic*, Stefan Panic

*Faculty of Technical Science, Kosovska Mitrovica, University of Pristina, 38 220 Kosovska Mitrovica, Kneza Milosa 7, Serbia*

ARTICLE INFO

*Keywords:*
Data mining
MATLAB
Linear regression
Classification tree
Discriminant analysis

ABSTRACT

Today's world depends largely on information and communication technologies. These technologies are in use in different areas of human life and work. Each day, more examples of possible application of information and communications technology are being discovered. In most cases, computer science is used to solve complex problems which have mathematical background. The most important and challenging job in agriculture is plant protection. This is due to its complexity and the lack of specialized tools that could predict when the conditions for specific infections are fulfilled. In this paper authors use different mathematics-based techniques for data processing and prediction of possible fruit disease infection. Six significant weather variables and one variable representing the month in the year are selected as predictor variables. Implemented techniques are compared with each other in order to select the best. Prediction includes two most important diseases of cherry fruit: *monilinia laxa* and *coccomyces hiemalis*. Data sets used in this research include data of eight year time period, collected over the region of Toplica in Serbia. The best achieved prediction accuracy is 95.8%. Additionally, the same implemented methods can be applied on other fruit species and other diseases for which data are known.

## 1. Introduction

In the past twenty years, along with increased fruit tree diseases infections, disaster areas has increased at least by 30% when compared with the stats recorded in the 90s. In recent years, pesticide is used mainly to control fruit tree diseases and insect pests and is used superfluously and abusively, leading to quality deterioration of fruit and endangering the human health. The naked eye observation of experts, and farmers experience are the main approaches adopted in practice for detection and identification of plant diseases (Weizheng et al., 2008). However, this requires continuous monitoring by experts which might be prohibitively expensive in large farms. Furthermore, in some developing countries, farmers may have to go long distances to contact experts, and this makes consulting experts too much expensive and time consuming (Babu and Srinivasa, 2010). There are two main requirements regarding fruit disease detection methods that must be fulfilled: speed and accuracy (Hahn et al., 2009). Fast and accurate prediction provides appropriate chemical fruit protection with pesticides that have preventive action (Arsevska et al., 2016). Such pesticides are cost-effective and efficient only if the application is performed at the right time. Additionally, with precise prediction farmers will reduce number of chemical treatments thus reducing residues in fruits. In most cases, successful prediction depends largely on the selection of predictable parameters on which to base. The most important and, in the same time, the least expensive for collection are weather parameters. Each disease can infect fruit tree if specific weather conditions are satisfied. If and only if the weather conditions are appropriate, disease infection will be possible. Predictable computer applications which could successfully predict when the weather conditions are favorable for disease development can greatly reduce the problems related to successful fruit protection.

In order to create appropriate prediction model, the influence of different prediction variables need to be considered. Prediction can be performed based on weather data and data that represent specific plant or disease characteristic. Some of the studies that support the use of data mining techniques (Chaurasia and Pal, 2017) for the plant disease prediction.

In paper (Georgiana, 2009) authors presented a small application of CART decision tree algorithm for weather prediction. Prediction is based on data collected in the period of four years. Selected prediction variables are: year, month, average pressure, relative humidity, clouds quantity, precipitation, and average temperature. From the aspect of appropriate conditions for disease infection, CART decision tree is also used as one of the models to forecast leaf wetness duration within the next 24 h, working with site-specific weather data estimates as inputs (Kim et al., 2006).

* Corresponding author.
*E-mail address:* srdjanjovic2016@hotmail.com (S. Jovic).

https://doi.org/10.1016/j.compag.2018.05.008
Received 19 March 2018; Received in revised form 15 April 2018; Accepted 6 May 2018

In the research presented in (Liu et al., 2005) authors create forty neural network prediction models that can be used according to weather factors from April to August. All models are created in order to predict the appearance of the eight kinds of diseases and insect pests on the fruit tree. Neural network prediction system is created using MATLAB neural network toolbox. Prediction models are based on four weather factors that are selected in every month as prediction parameters.

Authors in Rakesh et al. (2006) create an application based on SVM models for disease prediction in plant populations. Prediction is based on six significant weather variables that are selected as predictor variables. Application is created in MATLAB and used for rice blast prediction. Results show that SVM models combined with other statistical tools may offer flexible options for future forecasting models.

In Kim et al. (2005) fuzzy logic system is used for estimation of apparent infection rate by combining meteorological variables and biological criteria pertinent to soybean rust severity. A fuzzy logic apparent infection rate model is developed to simulate severity of soybean rust and validated using data from the field experiments on two soybean cultivars. In simulation of an entire epidemic period, the created model is able to predict disease severity accurately once initial values of disease severity are predicted correctly. Training data used to develop a daily apparent infection rate model are obtained from sequential planting experiments of soybean at the Asian Vegetable Research and Development Center, Taiwan. From the other side meteorological data are obtained from on-site weather station. Air temperature, dew period, and precipitation are selected to be prediction variables.

Prediction models for anthracnose disease of the tropical pasture legume *Stylosanthes scabra* prediction are presented in Chakraborty et al. (2004). Disease severity and weather data are analyzed by implementing artificial neural network models developed using data from some or all field sites in Australia and/or South America to predict severity at other sites. A rapid back-propagation learning algorithm is used in all prediction models and architectures with different numbers of hidden neurons explored to select the best architecture for the artificial neural network models. Results shows that artificial neural network models with weather parameters collected during the day and from the previous 24 h has the highest prediction success, and models trained on data from all sites over one continent correctly predicted disease severity in another continent in more than 75% of days (Rakesh et al., 2006). Moisture-related variables such as rain, leaf surface wetness and variables that impact on moisture availability such as radiation and wind are the most important weather variables in all artificial neural network models.

In the study presented in Lucky et al. (2016), authors examined the performance of multiple regression and three machine learning algorithms, namely: artificial neural networks, categorical and regression trees, and random forests, in predicting the pre-planting risk of *Stagonospora nodorum* blotch in wheat. Pre-planting factors tested as potential predictor variables are: cultivar resistance, latitude, longitude, previous crop, seeding rate, seed treatment, tillage type, and wheat residue. Disease severity assessed at the end of the growing season is used as the class variable. Results show that the random forests algorithm is the most accurate in predicting the risk of SNB, with an accuracy rate of 93%.

The most complex agricultural problem for all farmers and engineers is to predict when the appropriate conditions for disease infection are fulfilled. Proper and targeted observation and data processing could provide good information about dependencies among diseases and different weather and environmental parameters collected from the field. Huge amounts of data dating from the past could be treated by applying mathematical and statistical methods (Mohammadhassani et al., 2014; Toghroli et al., 2014; Toghroli et al., 2016; Mohammadhassani et al., 2013; Pan et al., 2016). These methods provide adequate support for detecting dependencies between data, outliers detection and prediction model creation.

Scope of this research includes data set creation and dependencies observation for two the most frequent cherry fruit diseases which can reduce the amount of fruit in the current production year and threaten the survival of the plant in the coming years. Research has three basic goals all of them indicating that even the agricultural problems could be solved using mathematical methods.

First goal pf the paper is data preprocessing, visualization, outlier detection, and reduction of redundancies among instances. Second goal is to create prediction models based on refined data after preprocessing by using different data mining prediction techniques. Third goal is comparison of prediction models, and selection of the best model. The degree of accuracy of the model is determined on the basis of statistical errors and data mining errors for the each model. The MATLAB (The MathWorks Inc., Natick, MA) program is used as the developing tool for all of the calculations and prediction model creation and testing.

## 2. Materials and methods

Data used in this paper are collected and organized particularly for the purpose of this research. This research is focused on two types of data sets. First data set type is the one containing data used for prediction model training. Second data set type represents data that are used for prediction model testing and accuracy evaluation.

Data from both datasets are organized in matrix form. The instances are stored in rows and each column corresponds to an attribute. Meteorological data used for creation of both data sets (training and test) are collected from the meteorological station placed at the College of Agriculture and food technology in Prokuplje (region Toplica in Republic of Serbia) that cover one hundred hectares of arable land. All meteorological data are measured three times a day – i.e. each measurement is performed for the time period of 8 h: the minimum, mean and maximum temperature, average humidity, rainfall and wind speed. The collected data are daily aggregated in the following way: the minimum temperature is calculated as minimum of three temperature minimums, the maximum is calculated as maximum of three temperature maximums and the averaging is performed on: temperature means, the average of humidity, and wind speed. The daily amount of rainfall is calculated as total sum of 8 h rainfall measurements.

The last column in the training data set represents the class attribute. Class attribute has predefined values that indicate the presence of the observed disease. History data about the presence of infections for the observed disease are collected from infections reports issued by certified institutions of Ministry of Agriculture, and they cover the same period as meteorological data (one value for each day). Based on this, class attribute values for both data sets are known. Class attribute has the following values: *nothing, both, monilinia and coccomyces*. First value (*nothing*) indicates that there are no appropriate conditions for disease infection, while the second value (*both*) means that there are conditions for the development of both diseases. Third and fourth values (*monilinia and coccomyces*) indicate one disease, specified by value itself. The forecasting of the class attribute in our study is performed using six numeric weather variables. Statistical analysis of numeric variables in the training data set is presented in Table 1.

Training data set contains 1224 instances. Data are collected in the period from 2009 to 2016, and each instance is obtained by

**Table 1**
Statistical analysis of numeric data values in training and test data set.

| Variable name | Min | Max | Mean | Variance | StdDev |
|---|---|---|---|---|---|
| Min temp | -2 | 28 | 12.97 | 24.52 | 4.95 |
| Max temp | 5 | 40 | 26.26 | 47.14 | 6.87 |
| Mean temp | 3 | 33 | 19.60 | 29.93 | 5.47 |
| Average humidity | 0.33 | 0.97 | 0.68 | 0.02 | 0.13 |
| Rainfall | 0 | 30 | 1.05 | 10.56 | 3.25 |
| Wind speed | 3 | 27 | 7.59 | 16.63 | 4.08 |