Computers and Electronics in Agriculture 154 (2018) 314–319

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

# Data mining based tool for early prediction of possible fruit pathogen infection

Bratislav Predic[a], Milos Ilic[b], Petar Spalevic[b], Slavisa Trajkovic[c], Srdjan Jovic[b,*], Andrija Stanic[b]

[a] Faculty of Electronic Engineering, University of Nis, 18000 Nis, Aleksandra Medvedeva 16, Serbia
[b] Faculty of Technical Science, University of Pristina, 38 220 Kosovska Mitrovica, Kneza Milosa 7, Serbia
[c] Faculty of Economics, University of Pristina, 38 220 Kosovska Mitrovica, Kolasinska 156, Serbia

ARTICLE INFO

*Keywords:*
Fruit pathogen infection
Classification
Data mining
Decision trees
Predictive models

ABSTRACT

Effective chemical protection of fruit is a complex task which can enable production of healthy food without chemical residues. As food health is a growing global concern, it is important to automate and optimize the fruit protection process. Different data mining techniques can be used to identify pattern diseases so as to prevent the excessive use of chemicals. However, application of data mining systems in this field is very complex task. Besides that, these systems are often designed for just one specific plant species. One solution for prediction a risk of fruit infection based on data that represent weather (meteorological) conditions and data pathogens is presented in this paper. The research is performed on the data collected at the region of Toplica in Republic of Serbia during five year period. In this paper data mining based tool for early prediction of fruit pathogen infection is performed. The application is based on the open source engine WEKA with GUI created in C#, and uses several data mining algorithms which are evaluated in this paper. Results shown that the prediction accuracy is 89%.

## 1. Introduction

Plant diseases can be responsible for decrease of the global food production therefore aggravating the problem of food shortages worldwide. The recent research shows that at least 800 million people suffer from malnutrition (Strange and Scott, 2005). The key challenge for all farmers and agricultural science is to find out how to prevent plant/fruit diseases and to perform an effective chemical treatment. This kind of problem requires close attention since wrong chemical treatments and the application of large amounts of chemicals could lead to soil contamination, and could endanger human health. Furthermore, chemical residues in human body could lead to serious illness. Fruit production agricultural science must find balance between successful chemical protection, and no fruit contamination. The key for solution of this problem is reduction of number of so called preventive chemical treatments. This problem can be avoided by using chemicals at the right time, when the smallest amount of chemicals can suppress the specific pathogen or pest. Main problem is how to predict the right time for suppression of pathogens. Different diseases that attack different fruit species cause infection in different conditions. These conditions are defined through laboratory experiments, and through observations carried out in the observed geographical area (Ivanovic, 1992; Stojanovic, 2004; Byrde and Willets, 1997; Peric, 2007).

The most important conditions for pathogen fruit infections are the appropriate weather conditions and presence of an active pathogenic spore. For the group of weather condition parameters we can consider temperature, amount of rainfall, humidity and leaf wetness. In these weather conditions an active pathogenic spore can produce infection. On the other hand, if the pathogenic spore is present in the air and on the fruit tree but is not active, the infection is not possible, regardless of weather conditions.

Data mining techniques can be used in agriculture to identify different patterns for many complex tasks. There are numerous studies that support the use of data mining techniques for weather prediction or prediction in the field of agriculture. Data mining techniques are often more powerful, flexible, and efficient for exploratory analysis than statistical techniques (Kantardzic, 2011). In the field of agriculture data mining is an emerging research method, and different data mining techniques are in use. Some of those techniques are artificial neural networks, decision trees, support vector machine, classification and clustering techniques. In study Hill et al. (2014) was used five machine learning algorithms (Decision Tree, Naive Bayes, Random Forest, AdaBoost, Support Vector Machine) and one statistical method (Logistic regression) (classifiers) to develop models to predict the outcome of

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

https://doi.org/10.1016/j.compag.2018.09.023
Received 7 August 2018; Received in revised form 9 September 2018; Accepted 16 September 2018

leafroller pest monitoring decisions on 'Hayward' kiwifruit crops in New Zealand. Predictions were made for spray and no-spray decisions based upon pre-determined acceptable rates of precision (proportion of correct decisions in test results). A real-time decision making system with which one can predict pests by means of a data mining technique and a wireless sensor network was developed in article Tripathy et al. (2011) where Gaussian Naïve Bayes and the Rapid Association Rule Mining were used. Based on these methods, they have conducted research on the prediction of when pests appearance based on various data collected from the wireless sensor network. In article Wang and Ma (2011) was compared Regression Method and SVM (Support Vector Machine) method to predict wheat stripe rust disease. This is one of the most critical plant diseases in China and prevention is very important. In the results of their experiment, the prediction using the SVM method showed high fitting and predictive precision, and was excellent on the side of speed. In research Hooker et al. (2002) was used MLR (Multiple Linear Regression) to determine periods and conditions of Deoxynivalenol occurrence. Their work collected relevant information from 399 farms in Ontario, southern Canada from 1996 to 2000 for the prediction of Deoxynivalenol occurrence. Daily precipitation, daily minimum and maximum temperatures, and relative humidity per hour were used as weather factors, and as a result of the experiment, it has been confirmed that time points of wheat growth, rainfall, and temperature are related with deoxynivalenol occurrence. In some of the papers authors used classification to classify soil based on morphological soil profile descriptions. They used clustering for plant and soil classification using colour images (Verheyen et al., 2001; Meyer et al., 2004). A procedure for early detection and differentiation of sugar beet diseases based on Support Vector Machine (SVM) and spectral vegetation indices was created in article Rumpf et al. (2010). The discrimination classification of healthy sugar beet leaves and infected leaves resulted in an accuracy of 97%. On the other hand, differentiation between healthy leaves and leaves with symptoms of three diseases had achieved accuracy higher than 86%. Degree of accuracy in both cases provides enough confidence in systems like these. In article Scherm (2014) was approached to the plant pathology domain from the Big Data perspective. In the world of ever-increasing volume of available data-sets they use meta-analysis in plant disease epidemiology and management among other fields. In article Ngugi et al. (2011) was quantified the relationships between different disease-related variables and its importance in plant disease prediction. Authors in articles Pietravalle et al. (2003) and Shah et al. (2013) demonstrated the importance of correlating weather data with plant disease occurrence. 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. 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, 2017) for the plant disease prediction. Based on the listed references there is relevance of relying on correlation of meteorological and plant disease data to predict and suggest adequate plant chemical treatment.

In this study software tool for an automatic prediction of the risk of fruit pathogenic is based on data mining techniques. Data mining is the field of discovering novel and potentially useful information from large amounts of data (Miamon and Rokach, 2010; Toghroli et al., 2018, 2016, 2014; Mansouri et al., 2017; Mohammadhassani et al., 2015; Safa et al., 2016). Data mining has two primary tasks, prediction and description. Predictive data mining task attempt to do forecast based on inference on available data. On the other hand, the description focuses on finding patterns describing the data that can be interpreted by humans. The used dataset in this study represents weather parameters and pathogen infection from the past. Data mining model is created based on the dataset. The main goals in this study are:

1) comparison of the achieved results (patterns) with those from references Ivanovic (1992), Stojanovic (2004), Byrde and Willets (1997) and Peric (2007) and
2) verification of the model against new data. This software tool could provide precise and timely prediction of the occurrence of plant diseases. Thus, farmers can be informed about the right moment for chemical protection and effectively produce healthier food treated with a minimum amount of required chemicals.

## 2. The software application

The implementation of one of the crucial components of the model that was proposed earlier (Ilic et al., 2015, 2018), and that is related to data pre-processing, processing and prediction is presented in this section. The implemented software operates on the data collected from the past, and predicts the occurrence of disease based on the training and test dataset.

### 2.1. Software architecture

The core functionality is implemented using open source data mining software called WEKA. WEKA or Waikato Environment for Knowledge Analysis software is product of the University of Waikato (New Zealand) and was first implemented in its modern form in 1997. It uses the GNU General Public License (GPL). The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results. It also has a general API, so WEKA can be embedded in other applications like any other library. The authors created WEKA .dll file and imported it in .NET environment. For this purpose the authors used IKVM.NET, which is an implementation of Java for .Net. It allows call of Java classes directly from .Net code and, via the GNU Classpath, provides most of the standard Java API for use in .Net.

This kind of implementation is chosen because WEKA classes provide high performance data mining capabilities and C# provides GUI and database connectivity. C# Windows Forms application is created as wrapper for WEKA classes in order to provide experimentation user interface for creation of dataset files, reviewing data in existing files and adding new data or removing older. Created application also contains support for all classification, clustering and prediction steps and reports generation. Because the weather data and data collected from spore traps is collected from historical records in paper form reports, it was necessary to allow end-users to put all data in dataset file trough intuitive and easy to use user interface. Also, the database support is important because the new reports may be collected from automated meteorological stations. In this way the future end-user can easily manipulate data and monitor classification or clustering process. Based on classification results, the user can create the best prediction model, start prediction and read the report on performed prediction.

### 2.2. Data collection and analysis

For our research and evaluation, we collected and prepared two data sets. The *training* dataset consists of data that covers period from 2011 to 2014. Because during the calendar year vegetation is active and diseases can infect fruit trees from April through August, only data from this period are used. This data is used for model training.

The test data set is created from data collected during 2015 and is used for testing of the created model. Meteorological data used for both data sets are collected from the meteorological station placed in Prokuplje (region Toplica in Republic of Serbia) that covers one hundred hectares of arable land. All meteorological parameters 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