Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 33 (2016) 332 – 339

Procedia Environmental Sciences

The 2nd International Symposium on LAPAN-IPB Satellite for Food Security and Environmental Monitoring 2015, LISAT-FSEM 2015

Web-based classification application for forest fire data using the shiny framework and the C5.0 algorithm

Gita Puspita Siknun*, Imas Sukaesih Sitanggang

Department of Computer Science, Bogor Agricultural University, Darmaga, Bogor 16680, Indonesia

Abstract

Forest fires are threats for our ecosystems and environment because their impact is very harmful. Every year, the number of hotspots increases, indicating the increase of forest fires in some regions in Indonesia, one of them is Riau Province. To predict the hotspot occurrence, we build a web-based application based on characteristics of area using the Shiny framework. We use the C5.0 algorithm by generating tree and rule-based classification models. The Shiny framework was implemented using reactivity expression, when an input changes, the server will rebuild the output based on the input data. We use the dataset of forest fires in Rokan Hilir district, Riau Province, in 2008. The dataset consists of ten explanatory layers (physical, weather, and socio-economic characteristics) and one target layer (hotspot or non-hotspot). The implementation of the C5.0 algorithm on forest fire data resulted tree models with accuracy of 72.72% and rule-based models with accuracy of 73.51%. The output of tree models is 16 classification rules while the output of rule-based models is 15 classification rules.

Peer-review under responsibility of the organizing committee of LISAT-FSEM2015

Keywords: C5.0; classification; forest fire; framework shiny; hotspot

1. Introduction

Forest fires are threats for our ecosystems and environment because their impact is very harmful. According to study in Bappenas and ADB on forest fires, total fire affected land area during El Nino (ENSO) event in 1997/1998

* Corresponding author. Tel.: +6285-6226-6650.
E-mail address: g.puspita27@gmail.com

Peer-review under responsibility of the organizing committee of LISAT-FSEM2015
doi:10.1016/j.proenv.2016.03.084

was about 9.75 million ha [1,2]. The fires emitted much carbon emissions and caused environmental impacts on regional to global scale.

The forest fire prevention is done by monitoring active hotspot data from satellite observations using geographic information system (GIS) to obtain the hotspot occurrences on some areas. Many hotspot occurrences found in logged or drained wetland and peat land areas in some districts in Sumatra, such as Rokan Hilir District, Riau Province. In 2014, there were 85 hotspots in Rokan Hilir District, scattered in several areas [3]. To reduce the incidence of forest fires, it is necessary to build an information system which is applying the data mining techniques. Data mining is able to process and analyze a large amount of data such as forest fire data.

Classification is one of data mining techniques. Previous research about developing classification models on spatial data has been conducted [4] by applying a spatial decision tree algorithm. In the study, the result of spatial decision tree was compared to ID3 and C4.5 algorithms resulted the classification models on non-spatial data. ID3 and C4.5 algorithms that were applied by some researchers have been available in desktop-based application, Weka data mining software. The spatial decision tree algorithm has resulted a higher accuracy of 71.66% than non-spatial classification models using the ID3 and C4.5 algorithm [4]. The ID3 algorithm had accuracy of 49.02% while the accuracy of C4.5 algorithm was 65.24%. Another research was conducted [5] to generate the classification rules using spatial entropy-based decision tree algorithm. The study used the 5-fold cross validation test to split both training set and testing set. However, the classification model on both researches was not implemented into an application that can be used by the related parties.

A recent study has been conducted to develop web-based applications. Hayardisi et al. [6] built a web-based OLAP (On-line Analytical Processing) integrated with a data warehouse for hotspot distribution data. The OLAP application provided the users summarization and visualization in crosstabs and graphs (bar plots and pie plots). Another study from Thariqa and Sitanggang [7] that developed web-based SOLAP (Spatial On-line Analytical Processing) by displaying the summary of hotspots based on socio-economic information. Based on the study, most of hotspots were occurred in a low population density and low school density. However, both studies have not developed an updating menu to handle new hotspot data and the users need to predict the hotspot occurrences based on another characteristics.

A web-based geographic information system (GIS) was developed using OpenGeo Suite for classifying hotspot occurrences [8]. The research applied the C4.5 algorithm in a Java implementation called J48 using Weka data mining software. The application has a menu to predict the hotspot occurrences based on area characteristics with 69.56% accuracy. However, the area characteristics only consisted of the location of the distance to nearest main cities, roads, and rivers. Moreover, if other supporting factors are added in the GIS, the dataset should be re-classify using Weka and the classification rules should be re-implemented in the GIS.

A web-based application using the Shiny Framework was developed using the DBSCAN algorithm [9]. The study performed the clustering method on two dataset of hotspots on Kalimantan Island and South Sumatera Province in 2002-2003. The application displayed the spread pattern of hotspots from the chosen data by user. However it has not implemented a menu to upload a new dataset.

In 1980, JR Quinlan introduced the C4.5 algorithm as the development of ID3 algorithm. Then JR Quinlan was continually developing the classification tree and rule-based models into C5.0 algorithm [10]. C5.0 was proprietary and commercially available until 2011 when a GPL version was released [10].

Previous researches have not implemented the classification models into a web-based application that can be accessed by the related parties to prevent forest fires. Hence, this study developed a web-based application to display classification models and to predict hotspots occurrence. The Shiny framework can build a web-based application which is published on the Internet. It is easily developed and integrated with a web content using HTML and CSS [11]. The Shiny Framework is an R package which is responsive to display the results of the data mining analysis into web-based applications. This web-based application was developed using the R programming language which is open source with several popular packages [12]. In the R package, there are many classification algorithms that can be easily used by users. In this study, the classification algorithm applied is the C5.0 algorithm.

2. Data and methods

2.1. Study area and data

The data used in this study is forest fire dataset in Rokan Hilir District, Riau Province, in 2008. The dataset was obtained and described in Sitanggang et al [4]. The dataset consists of ten explanatory layers (distance to the nearest city centers, distance to the nearest rivers, distance to the nearest roads, income sources, land cover, peat land types, peatland depth, screen temperature, wind speed, and precipitation) and one target layer (hotspot or non-hotspot points).

The hotspot distributions and coordinates of burn area obtained from FIRMS MODIS Fire/Hotspot, NASA/ University of Maryland. The non-hotspot points were obtained by generating random points outside a hotspot buffer [13]. The dataset in physical characteristics are the distance to the nearest roads, rivers, and city centers, land cover, administrative boundaries provided in digital map. They were obtained from Indonesian Geospatial Information Agency (BIG). The other dataset for peat land depth and peat land types were obtained from wetland International in digital map. The socio-economic data consists of income sources which were obtained from Statistics Indonesia (BPS). The weather data consists of precipitation in mm/day, wind speed in m/s, and screen temperature in K obtained from Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG). They were provided in NetCDF format. The dataset had been pre-processed and saved as CSV format.

2.2. C5.0 algorithm

Classification is the task of learning a target function f that maps each attribute set x to one of the predefined class label y [14]. The target function is also known as a classification model that can also be used to predict the class label of unknown records [14]. The model classification used is a tree model and a rule-based model. A tree model has a flowchart-like tree structure, where each internal node (non-leaf node) indicates the test on an attribute, each branch represents the results of the test, and each leaf node is a class label [15]. A rule-based model is a set of if-then conditions that have been derived from a tree model into more simple conditions. In this study, the tree and rule-based models implemented C5.0 algorithm. The C5.0 algorithm is generated based on decision trees [16]. It can derive a set of if-then rules, shows the easier and more interpretable rules to understand. The C5.0 decision tree algorithm can be seen in Fig. 1 [15]. The C5.0 algorithm was the development of C4.5 algorithm as an improved version and it has several advantages [17]. The C5.0 trees and rule sets are smaller than the C4.5 counterparts. The trees are constructed using recursive manner (divide and conquer) from a set of training cases [18].

2.3. Shiny framework

Shiny is a framework from R and one of package [19] we used to build an interactive web application. Before displaying the classification models on the web page, we used the C50 package [20] for generating the models and the Caret package [21] for splitting the data. A random proportion of data were splatted into 75% to train a model while the remainder was used for prediction. Then we applied the models to display them on the web page and implemented the models to predict the hotspot occurrences based on characteristics of area using the framework Shiny.

Algorithm: Generate_decision_tree
Input:
a. Data partition, *D*, a set of training tuples and their associated class labels
b. *attribute_list, the set of candidate attributes*
c. *attribute_selection_method*, a procedure to determine the splitting criterion that best partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and, either a *split-point* or *splitting subset*

Output: a decision tree
Method:
1. create a node *N*
2. if tuples in *D* are all of the same class, *C, then*
3. return *N* as a leaf node labeled with the class *C*
4. if *attribute_list is* empty, then
5. return *N* as a leaf node labeled with the majority class in *D*
6. apply *attribute_selection_method* (*D, attribute_list*) to find the best *splitting_criterion*
7. label node *N* with *splitting_criterion*
8. if *splitting_attribute is discrete-valued and multiway splits allowed then*
9. *attribute_list* ← *attribute_list – splitting_attribute*
10. for each outcome *j* of *splitting_criterion*
    i. let *Dj* be the set of data tuples in *D* satisfying outcome *j*
    ii. if *Di* is empty then
        attach a leaf labeled with majority class in *D* to node *N*
    iii. else, attach the node returned by Generate_decision_tree (*Dj*, *attribute_list*) to node *N*
11. return *N*

Fig. 1. C5.0 algorithm for inducing a decision tree from training tuples.

3. Results

3.1. Data classification using C5.0 algorithm

Applying the C5.0 algorithm on the forest fire dataset results the number of rules. There are 16 rules generated from the tree model and 15 rules generated from the rule-based model. Several rules from tree model can be seen in Figure 2 while several rules from rule-based model can be seen in Fig. 3. Accuracy of the tree models on the testing set is 72.72% meaning that 69 of 253 hotspot points are incorrectly classified by the tree while accuracy of the rule-based models is 73.51% meaning that 67 of 253 hotspot points are incorrectly classified.

Each rule has a statistics (n/m) and a lift on the rule-based model. To a leaf, n is the number of training cases covered by rules and m, shows how many of them do not belong to the class predicted by the rule. In the 3rd rule, the statistics (165/27) is classified as a class of 165 hotspot points and a class of 27 non hotspot points (Fig. 3). The lift is the result of dividing the Laplace ratio by the relative frequency of the predicted class in the training set. Based on the classification results, both of tree model and rule-based model have different usage variables on the generated rules.

The usage variables show the list of control parameters for predicting the class, hotspot. Both of tree model and rule-based model have the same variables with different percentages and each of them are specified in Table 1. Peat land type are the highest usage variable in both of the models with 100% usage in tree models shows that each of rules generated have peat land type while 93.16% usage in rule-based models specify some rules have it.

1. IF peatland_type in {(Hemists/Saprists(60/40), Moderate), (Saprists/min(50/50), Moderate), (Hemists/min(30/70), Moderate), (non_peatland), (Saprists/min(50/50), Shallow), (Saprists/Hemists(60/40), Moderate)} AND dist_road in {(Low: <= 2.5 km), (High: > 5 km)} AND dist_river in {(Low: <= 1.5 km), (Medium: (1.5 km, 3])} THEN Class Non Hotspot (214/41)
2. IF peatland_type in {(Hemists/Saprists(60/40), Very_deep), (Saprists/min(90/10), Moderate), (Hemists/Saprists(60/40), Deep), (Saprists(100), Moderate)} AND income_source in {(Others), (Trading_restaurant), (Agriculture)} THEN Class Non Hotspot (18/5)

Fig. 2. Classification rules from tree model.

1. IF dist_river in {(Low: <= 1.5 km), (Medium: (1.5 km, 3])} AND dist_road in {(Low: <= 2.5 km), (High: > 5 km)} AND peatland_type in {(Hemists/Saprists(60/40), Moderate), (Saprists/Hemists(60/40), Moderate), (Saprists/min(50/50), Moderate), (Hemists/min(30/70), Moderate), (non_peatland), (Saprists/min(50/50), Shallow)} THEN Class Non Hotspot (214/41, lift 1.6)
2. IF income_source in {(Others), (Trading_restaurant), (Agriculture)} THEN Class Non Hotspot (138/35, lift 1.5)
3. IF income_source in {(Medium: (2.5 km, 5 km]), (Other_agriculture), (Forestry)} AND peatland_type in {(Saprists/min(90/10), Moderate), (Hemists/Saprists(60/40), Deep), (Saprists(100), Moderate)} THEN Class Hotspot (165/27, lift 1.6)

Fig. 3. Classification rules from rule-based model.

Table 1. Usage variables on tree model and rule-based model.

Usage variables	
Tree model	Rule-based model
100% Peatland type	93.16% Peatland type
97.63% Road distance	82.50% Road distance
54.61% River distance	63.95% Income source
51.71% Income source	42.50% River distance
10.39% Land cover	33.16% Land cover
10.00% Wind speed	31.58% Wind speed

3.2. Web-based application using Shiny

Web-based application was built using the Shiny package [19] with Rstudio. The Shiny framework has two structures, contain a server.R file and ui.R file. The server file is a set of instructions that build the R components while the user-interface file is a set of instructions to display the application.

On the main page of the application, users upload the forest fire data as a csv file. The data are received by server as a data frame (Fig. 4). The data frame, built by reactive function, is stored in datasetInput function. The datasetInput function is used as an object to build the partition data using the Caret package [20] and to obtain the classification models using the C50 package [21].

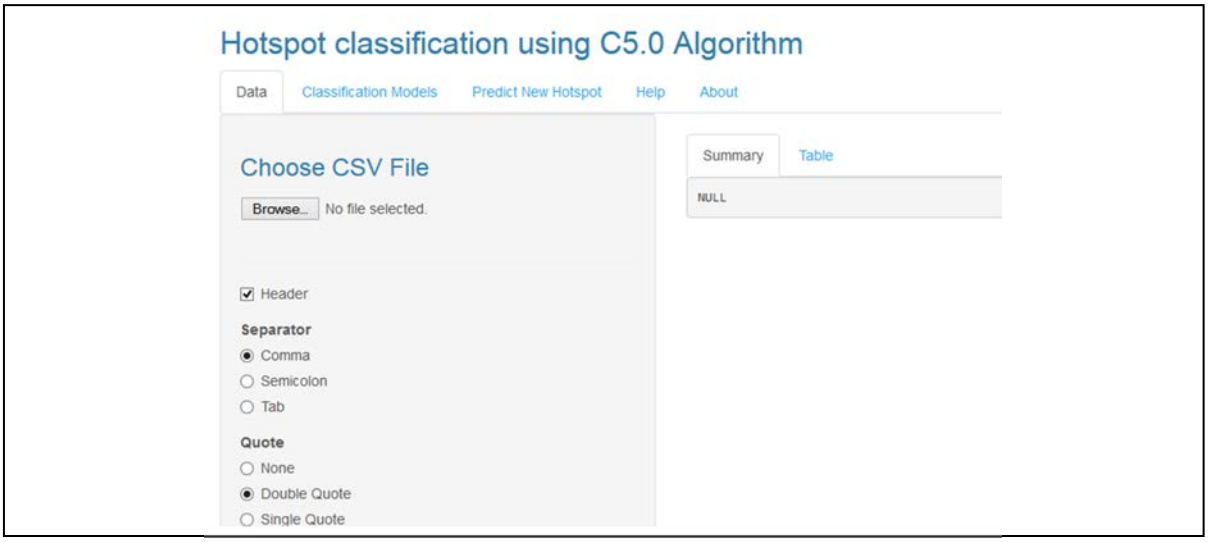

Fig. 4. View of web-based classification main page.

The classification models that have been generated can be used to predict the hotspot occurrence on the Predict New Hotspot tabPanel (Fig. 5). The prediction function depends on classification function that has been implemented into the model reactively. When an input of forest fire dataset changes, the classification models will rebuild and the prediction function will automatically update in its changes. The prediction function is in the C50 package and it can generate the model confidence values or predicted class.

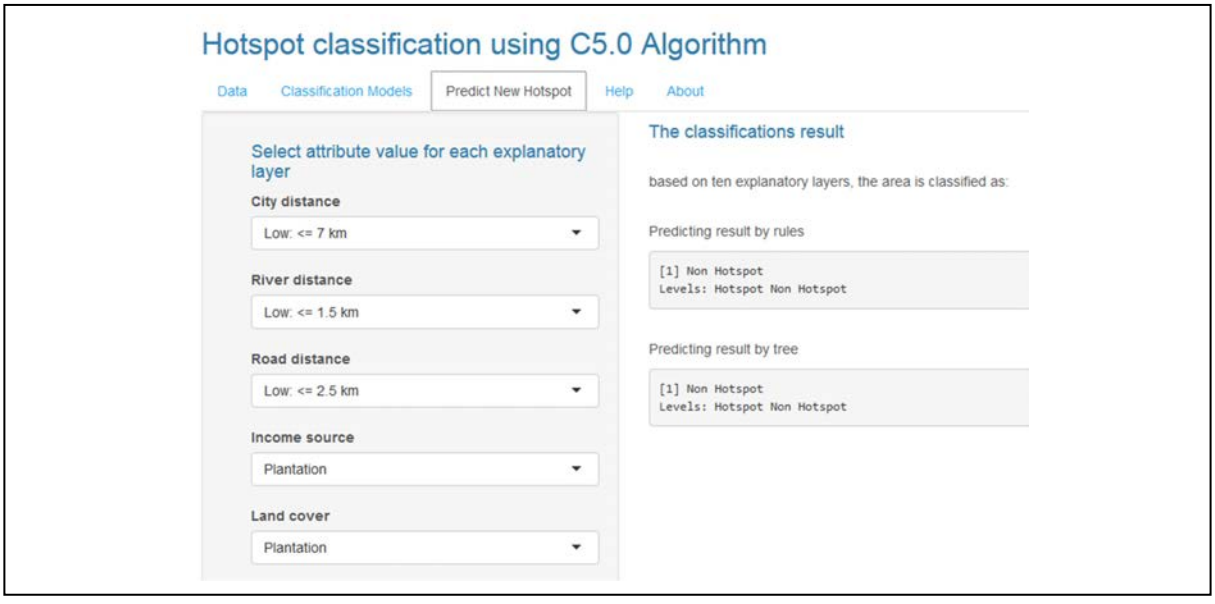

Fig. 5. View of prediction result on predict new hotspot tabPanel.

3.3. Hotspot prediction

On Predict New Hotspot tabPanel, a new forest fire data was tested to obtain a prediction result. This study predicted the hotspot occurrence for area characteristics with peatland type is Hemists/ Saprists(60/40), Deep), the nearest distance to the road is 3 km, and income source is forestry. The nearest distance to the road (3 km) is categorized as medium in this application. This application displays the prediction result based on the area characteristics as Hotspot point (Fig. 6).

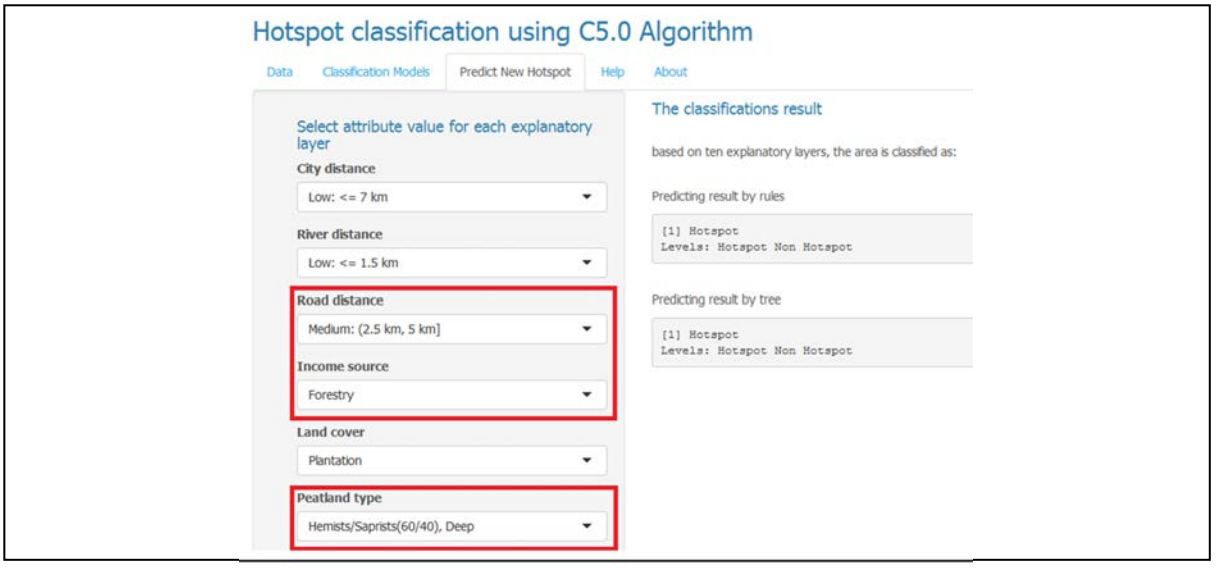

Fig. 6. Prediction result for hotspot occurrence as hotspot point.

Our future work will focus on several improvements including 1) implementing other algorithm on the same study to get the classification models with higher accuracy; 2) expanding this work with other data mining application for hotspot data analysis; 3) preparing bigger dataset for other study area, like in Sumatera and Kalimantan Province.

4. Conclusion

This research applied the C5.0 algorithm on the forest fire dataset and has built a web-based application using the Shiny framework. The fire forest dataset consists of physical, weather, and socio-economic that influence hotspots occurrence in the study area, Rokan Hilir District, Indonesia. The accuracy of classification models for predicting hotspots occurrences are 72.72% on tree models and 73.51% on rule-based models. The number of rules generated from the tree models is 16 while from the rule-based models is 15. Based on the usage variables, peat land type, road distance, river distance, income source, land cover, and wind speed influence the hotspots occurrence. The tree and rule-based models have been successfully implemented for predicting the hotspots occurrence based on the area characteristics. On Shiny framework, the reactivity function has been implemented to create each output based on the input changes by users.

References

1. Bappenas, ADB. *Causes, extent, impact and costs of 1997 / 1998 fires and drought*. Jakarta, ID: Bappenas and ADB; 2003.
2. Tacconi L. *Fires in Indonesia: causes, costs and policy implications*. Bogor: Center for International Forestry Research (CIFOR); 2003.
3. Ali M. The number of hotspots in Riau increased to 130. 2014 [cited 2014 October 25]; Available from: http://news.liputan6.com.

4. Sitanggang IS, Yaakob R, Mustapha N, Ainuddin AN. A decision tree based on spatial relationships for predicting hotspots in peatlands. TELKOMNIKA 2014; **12**: p. 511-518.
5. Nurpratami ID, Sitanggang IS. Classification rules for hotspot occurrences using spatial entropy-based decision tree algorithm. Procedia Environmental Sciences 2015; **24**: p. 120-126.
6. Hayardisi G, Sitanggang IS, Syaufina L. Data warehouse and web-based OLAP for hotspot distribution in Indonesia. 2nd Conference on Data Mining & Optimization; 2009.
7. Thariqa P, Sitanggang IS. Spatial online analytical processing for hotspots distribution based on socio-economic factors in Riau Province Indonesia. *Procedia Environmental Sciences* 2015; **24**:277-84.
8. Amri K, Sitanggang IS. A geographic information system for hotspot occurrences classification in Riau Province Indonesia. Procedia Environmental Sciences 2015; **24**:127-31.
9. Nisa KK, Andrianto HA, Mardhiyyah R. Hotspot clustering using DBSCAN algorithm and Shiny web framework. *IEEE* 2014; 129-32.
10. Kuhn M. *Classification using C5.0 UseR! 2013*. Groton CT: Pfizer Global R&D; 2013.
11. Venables WN, Smith DM. An Introduction to R; 2014 [cited 2014 October 24]; Available from: http://cran.r-project.org/doc/manuals/r-release/R-intro.html.
12. Beeley C. *Web application development with R using Shiny*. Birmingham: Packt; 2013.
13. Sitanggang IS, Yaakob R, Mustapha N, Ainuddin AN. Burn area processing to generate false alarm data for hotspot prediction models. *TELKOMNIKA* 2015; **13**:1037-46.
14. Tan PN, Steinbach M, Kumar V. Introduction to Data Mining. Boston. US: Pearson Addison Wesley; 2005.
15. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. 3rd ed. Massachusetts, US: Morgan Kaufmann; 2012.
16. Information on See5/C5.0-RuleQuest Research Data Mining Tools. 2011 [cited 2015 January 10]; Available from: http://www.rulequest.com/see5-info.html.
17. Is See5/C5 Better Than C4.5?. 2009 [cited 2015 January 10]; Available from: http://www.rulequest.com/see5-comparison.html.
18. Quinlan JR. *C4.5: programs for machine learning. vol. 1*. California: Morgan Kaufmann; 1993.
19. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J, Otto M. Package 'shiny'. 2015 [cited 2015 January 12]; Available from: http://shiny.rstudio.com.
20. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A. Package 'caret'. 2015 [cited 2015 February 4]; Available from: http://caret.r-forge.r-project.org.
21. Kuhn M, Weston S, Coulter N, Culp M. Package 'C50'. 2015 [cited 2015 February 4]; Available from: http://cran.r-project.org/web/packages/C50/index.html.