Journal of Engineering and Applied Sciences 13 (4): 903-909, 2018
ISSN: 1816-949X

Effective Prediction Model for Colorectal Cancer using Decision Tree and Clustering

Yong Gyu Jung and Min Soo Kang
Department of Medical IT, Eulji University, 553 Sanseong-daero, Sujeong-gu,
Seongnam-si, 13135 Gyeonggi-do, Korea

Abstract: Recently, data mining techniques have been utilized in various fields in order to effectively classify and analyze the information desired by a large number of data. Decision trees and clustering are basic and widely used techniques. We will use decision trees and clustering to analyze colon cancer and predict effective treatment. So, we will analyze the decision tree and clustering in detail to predict effective treatment and then study better treatment. In order to use various algorithms such as J48 decision trees and to selectively use the optimal algorithm for a specific situation, the fitness of the algorithm for each situation. In this study, we evaluated the performance of the algorithm based on the results of 10 cross validation using decision trees and explained the treatment prediction of colon cancer based on the result of clustering. We analyzed the performance of the algorithm and the results of the experimental data.

Key words: Colorectal Cancer (CRC), bowel cancer, decision tree, clustering, J48, situation, treatment

INTRODUCTION

Colorectal Cancer (CRC), also known as bowel cancer and colon cancer is the development of cancer from the colon or rectum (parts of the large intestine). A cancer is the abnormal growth of cells that have the ability to invade or spread to other parts of the body. Signs and symptoms may include blood in the stool, a change in bowel movements, weight loss and feeling tired all the time. Most colorectal cancers are due to old age and lifestyle factors with only a small number of cases due to underlying genetic disorders. Some risk factors include diet, obesity, smoking and lack of physical activity. Dietary factors that increase the risk include red and processed meat as well as alcohol. Another risk factor is inflammatory bowel diseasewhich includes Crohn's disease and ulcerative colitis. Some of the inherited genetic disorders that can cause colorectal cancer include familial adenomatous polyposis and hereditary non-polyposis colon cancer, however, these represent <5% of cases. It typically starts as a benign tumor, often in the form of a polypwhich over time becomes cancerous (Witten *et al*., 2011; Ruiz, 2009).

Bowel cancer may be diagnosed by obtaining a sample of the colon during a sigmoidoscopy or colonoscopy. This is then followed by medical imaging to determine if the disease has spread. Screening is effective for preventing and decreasing deaths from colorectal cancer. Screening, by one of a number of methods is recommended starting from the age of 50-75. During colonoscopy, small polyps may be removed if found.If a large polyp or tumor is found, a biopsy may be performed to check if it is cancerous. Aspirin and other non-steroidal anti-inflammatory drugs decrease the risk. Their general use is not recommended for this purpose, however, due to side effects (Lamma *et al*., 2006; Davis *et al*., 2004).

Treatments used for colorectal cancer may include some combination of surgery, radiation therapy, chemotherapy and targeted therapy. Cancers that are confined within the wall of the colon may be curable with surgery while cancer that has spread widely are usually not curable with management being directed towards improving quality of life and symptoms. The 5 years survival rate in the United States is around 65%. The individual likelihood of survival depends on how advanced the cancer is whether or not all the cancer can be removed with surgery and the person's overall health. Globally, colorectal cancer is the third most common type of cancer, making up about 10% of all cases. It is more common in developed countries where more than 65% of cases are found. It is less common in women than men.

In this study, data clustering will be performed using the EM algorithm among the clustering techniques. EM starts with estimating the initial values for several parameters and uses this parameter to calculate the probability that each piece of data belongs to a cluster. Next, we re-estimate the parameters using the calculated probabilities and repeat this process. We also use

Corresponding Author: Min Soo Kang, Department of Medical IT, Eulji University, 553 Sanseong-daero, Sujeong-gu, Seongnam-si, 13135 Gyeonggi-do, Korea

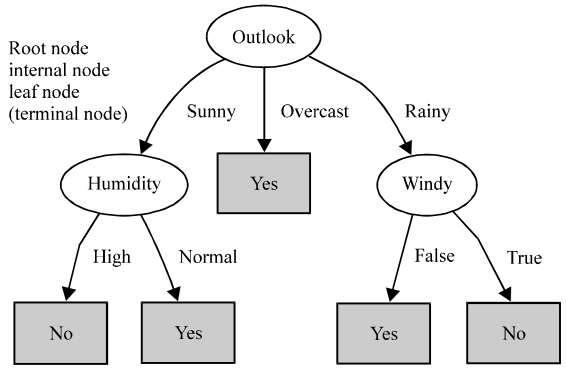

Fig. 1: Decision tree example

'attribute selection' as a filterwhich is a filter that allows you to select the attributes of the data and view them separately. In this study, we will randomly select attributes from attribute selection using random search.

Theory

Decision trees: Decision trees are widely used in various fields as a typical analysis method of data mining analysis. It is used to classify for prediction of the target variable. To apply the target variable to numeric data, the target variable be discretized from a numeric variable into categorical variable (Fig. 1). We will analyze the data with the J48 widely known to the public among the decision tree learners. It is one of the special algorithms used for predicting the data. Decision trees have several advantages. One is to express the results of data analysis through decision trees in a tree structure and it is easy for users to understand and explain the results. In addition, the accuracy of the classification rate is lower than that of classification methods such as neural network or logistic regression analysis. However, it is often used because it can be easily understood and explained and can be directly used for decision making. When the characteristics of the data could not be divided vertically and horizontally into specific variables, the classification rate is lowered and the tree becomes complicated. This is because different algorithms consider several variables simultaneously but the decision tree selects only one variable. The tree shape can vary greatly depending on the difference and the configuration can be different even when two variables have similar information (Moran *et al*., 2009; Han and Kamber, 2001).

Clustering: Clustering is an analytical technique in which individuals or objects are grouped into several clusters, so that, individuals having similar characteristics by similarity or distance are grouped together. The main purpose of data clustering is to identify the characteristics of each clustered population. In particular, unlike

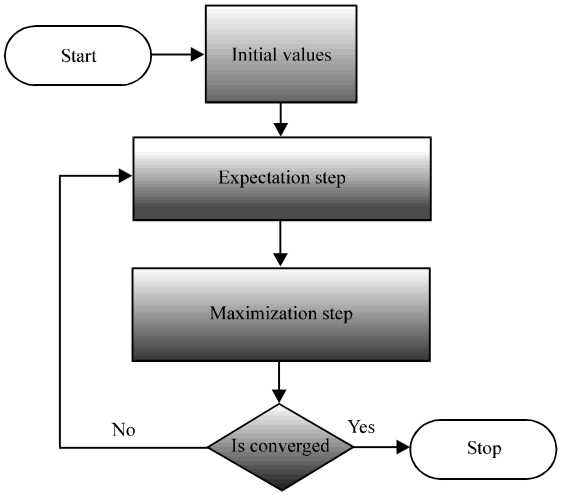

Fig. 2: EM algorithm example

classification schemes (such as decision trees), there is no assumption about the number of groups or the structure of groups, only a cluster is formed by similarities or distances between individuals this is a technique that can be utilized in a situation where there is no clear classification standard or unknown.

Figure 2 is similar to the clustering algorithm. The procedure of the EM algorithmwhich extends to an unlabeled large data set, first trains the classifier using the labeled data. Next, after classifying the classifier with unlabeled data, apply it to the unlabeled data and label them with class probabilities ('E: Expectation' step. Third, label all data ('M: Maximization' stage) and repeat this action until convergence. Finally, we will cluster and analyze the colon cancer data and predict the treatment and prevention methods. The k-means algorithm is also a typical algorithm of clustering and we will see how it differs from the EM used in this study. It shows the clustering of Iris data using the k-means algorithm and it shows the clustering of Iris data using the EM algorithm. The above data is divided into similar shapes using clusters of k-means and EM. However, there is a slight difference when analyzing it and cheating it. In the results of the k-means analysis, a cluster is formed in the shape of a circle because the cluster is assigned to the minimum distance at the center point. However, in the EM analysis result, it is assigned to the distribution in the shape of an ellipse rather than a circular shape in order to form a two-dimensional normal distribution (Hastie *et al*., 2001).

MATERIALS AND METHODS

Experiments

Experimental data: Colon cancer refers to a malignant tumor consisting of cancer cells in the large intestine.

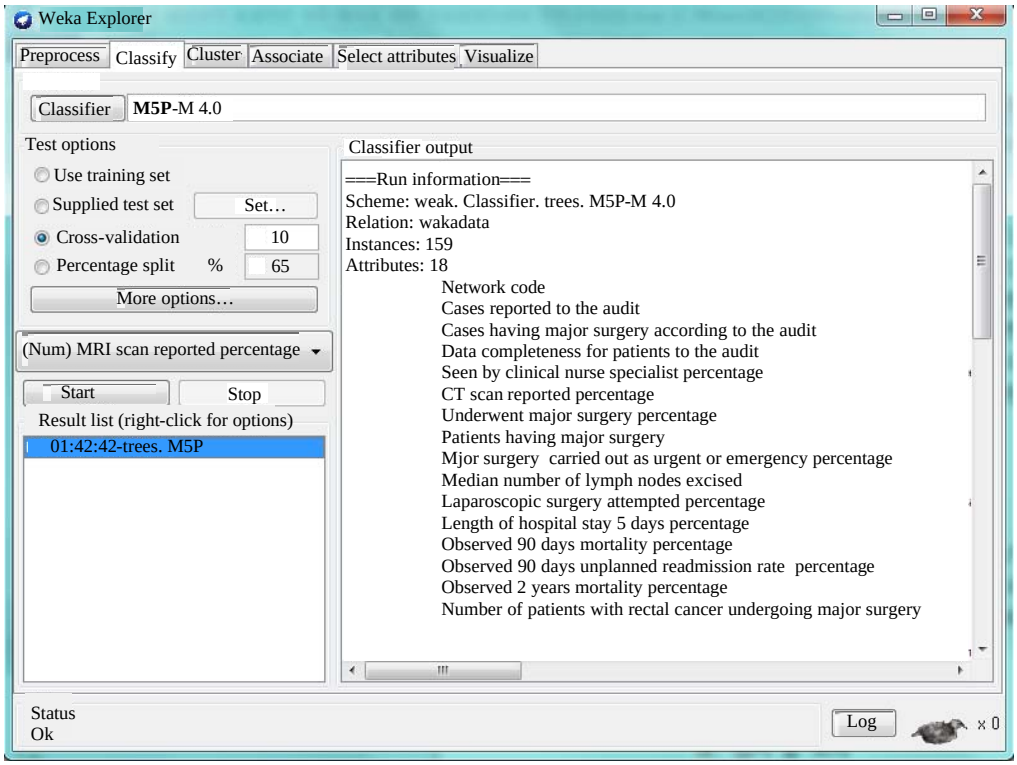

Fig. 3: Data analysis using decision tree M5P algorithm

There are many types of colon cancer including colon cancer and rectal cancerwhich are collectively referred to as colon cancer or colorectal cancer. Cancer of the leg are different but known by approximately 25%, so, mutually beneficial colon, descending colon missed 15, 5%, s colon 25, 10% scolon rectal junction, rectum 20% per person. The causes of colorectal cancer can be divided into environmental factors and genetic factors and the incidence varies depending on the regional characteristics. High calorie intake, animal fat intake and lack of fiber intake are associated with the development of colorectal cancer (Ruiz, 2009). The most notable cause of colorectal cancer among them is the excessive consumption of meat or high fat diet but high intake of animal fat with high saturated fat content increases the chance of getting colon cancer. Another cause is lack of fiber intake. Calcium, lack of vitamin D, bake or fry cooking, lack of exercise, inflammatory bowel disease, colon polyps, genetic factors and many other things can cause colon cancer (Fig. 3).

Symptoms of colorectal cancer initially do not show any symptoms. So, it is difficult for many people to find it early and it is only after a while that they can see if the disease has been sick. Chapter is the blood loss as bleeding appear a strain introduced or or anorexia, weight loss and if advanced cancer has also appeared a change in bowel habits such as stomachache or diarrhea or constipation caused of rectal bleeding blood from the anus symptoms may appear. However, most of the symptoms are not so, distinctive and often cited as the end of colon cancer even know the disease. Periodic health screening is therefore, necessary.

As a preventive measure for colorectal cancer it is good to check your physical condition once a periodic health checkup. This can be a preventive measure for all diseases as well as colorectal cancer. And the habit of checking bowel movements such as the color or shape of the stomach can be a good habit to find colon cancer. And while life appears anorexia and weight loss as appears a change in bowel habits such as diarrhea, constipation, pain begins and the embryonic bleeding from the anus and the lump touched the ship should suspected colonic polyps and colorectal cancer, colorectal polyps. In the case of adults aged 40 years or older, the incidence of colorectal cancer is higher. If colonoscopy is performed every 1-2 years with colonoscopy, colon cancer will be detected early and colon cancer survival rate will increase. And frequent exercise and dietary fiber-rich vegetables and fruits can help prevent colon cancer.

Table 1: Experiment data attributes

Network trust code/ Network/trust name	No. of cases reproted to the audit	No. of cases identified in HES	Case ascertainment (%)	No. case having major surgery according to the audit	Data completeness for patients having major surgery (%)	No. of patients reported to the audit	Discussed at MDT meeting (%)
N52							
Northern England SCN						1935	99.4
RLN							
City Hospitals Sunderland NHS Foundation Trust	173	195	89	109	86	173	98.3
RXP							
Durham and Darlington NHS Foundation Trust	299	287	104	205	93	299	100
RR7							
Gateshead Health NHS Foundation trust	145	157	92	106	93	145	100
RNL							
North Cumbria University Hospitals NHS Trust	201	235	86	108	79	201	98
RVW							
North Tees and Hartiepool NHS Foundation Trust	235	257	91	155	92	235	100

Network trust code/ Network/trust name	Seen by clinical nurse specialist (%)	CT scan reported (%)	Underwent major surgery (%)	No. of patients having major surgery (%)	Patients with distant met astases	Major surgery carried out urgent or emergency (%)
N52						
Northern England SCN	95.5	97.4	64.8	1253	9.3	13.1
RLN						
City Hospitals Sunderland NHS Foundation Trust	92.9	97.7	63	109	9.7	9.2
RXP						
Durham and Darlington NHS Foundation Trust	100	99	68.6	205	6.5	11.7
RR7						
Gateshead Health NHS Foundation Trust	97.8	97.2	73.1	106	10.9	10.4
RNL						
North Cumbria University Hospitals NHS Trust	87.4	97	53.7	108	10	12.1
RVW						
North Tees and Hartiepool NHS Foundation Trust	92.9	98.7	66	155	6	12.3

Property variables (data attributes): These experimental data have various numerical properties as attribute values related to colorectal surgery. It is composed of various numeric attributes and the name of the patient network that underwent this colorectal surgery. The details of each attribute are shown in Table 1.

First, the network code and name are defined as nominal attributes and various other attributes are described as numerical attributes. We will analyze the number of patients undergoing major surgery and the number of patients undergoing various tests such as CT and MRI to determine whether colon cancer is

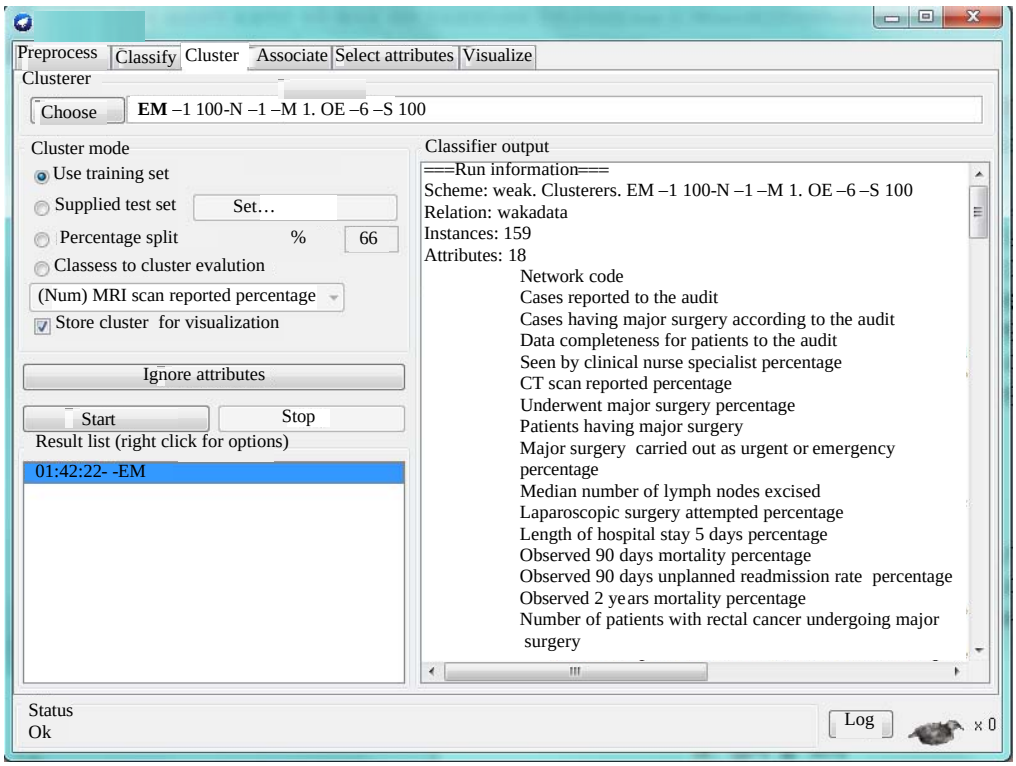

Fig. 4: Data analysis using clustering EM algorithm

being treated well or whether treatment prevention predictions will be made through experimental tools.

Experimental process: Experimental data describes colorectal surgery and what the patient has undergone. The number of patients who underwent major surgery and the mortality rate, the number of people who have been examined by CT, MRI, etc. and the various attributes and their attribute values are shown. The larger the number of surgeries, the more surgeries or surveys are performed. The data used are data on colorectal cancer and data on the number of operations and the number of surgeries of people with colorectal cancer (Fig. 4).

RESULTS AND DISCUSSION

The experiment was performed using decision tree and clustering using colorectal cancer data. First, we analyzed the data using M5P algorithm of decision tree. For the data analysis method, the fold value of cross-validation was set to 10. This is a ten-fold repetition of evaluating the 10-point cross-validation. Algorithm 1 shows the results of M5P in decision tree algorithm for colorectal cancer data. Colorectal cancer data consisted of one Linear Model. In the summary at the bottom of the Algorithm 1 the correlation coefficient shows 0.1355, the mean absolute error is 5.5067, the root mean squared error is 7.128, the relative absolute error (root mean square error) relative absolute error) was 96.8756% and root relative squared error was 98.7206%.

Algorithm 1: Result using decision tree via. M5P

Classifier Model (full training set)
M5 Pruned Model tree:
(using smoothing linear model)
LM1 (59/9.447%)
LM num: 1
MRI scan reported percentage =
7.5844*Network code = RXW, RLN RRI, RNA, RMC, RXL, RXN, RGG, RJL, RBT, RNL, RTF,1
+3.3589*Network code = RXL, RXN, RFF, RJL, RBT, RNL, RFT, RP5, RXR, RCD, N56, N5
+2.687*Network code = RJL, RBT, RNL, RIF, RP5, RXR, RCD, N56, N53, RWA, RTX, RVY
0.8101*Network code = RP5, RXR, RCD, N56, N53, RWA, RTX, RVY, RW3, RJC, RMP, N5
+0.9784*Network code = N53, RWA, RTX, RVY, RW3, RJC, RMP, N51, RWJ, RO6, RHO, RL
+1.1513*Network code = RTX, RVY, RW3, RJC, RMP, N51, RWJ, RO6,RL4, RJR, N5
+0.6941*Network code = RMP, N51, RWJ, RO6, RHO, RL4, RJR,N50, RXK, RCBCA, RBN
+1.6416*Network code = RWJ, RO6, RHO, RL4, RJR, N50, RXK, RCBCA, RBN, N52, RE9
+1.5313*Network code = RXK, RCBCA, RBN, N52, RE9,

RM2, RJN, RBV, RWW, RBL, RR7
+1.1606*Network code = RJN, RBV, RWW, RBL, RR7, RCF, RTD, RWY, RW6, RTR, RXP, RF
+2.2315*Network code = RTD, RWY, RW6, RTR, RXP, RFS, RXF, RCB55, REM, RLT, RVW
+1.1197*Network code = RTR, RXP, RFS, RXF, RCB55, REM, RLT, RVW, RM3, RRF, RFR
+3.5847*Network code = REM, RLT, RVW, RM3, RRF, RFR
-0.0026 case reported to the Audit
+71.9167

Number of rules:1
Time taken to build model : 0.03sec
=== Cross -validation ===
=== Summary ===

Correlation coefficient	0.1355
Mean absolut error	5.5067
Root mean square error	7.128
Relative absolute error	96.8756
Root relative square error	98.7206
Total number of instances	59
Ignored class unknown instances	100

Evaluation of results: It shows the experimental results of the 10 fold cross validation by applying the algorithm M5P of the decision tree and the algorithm EM of the clustering, respectively. Colorectal cancer data are divided into five clusterswhich are unevenly distributed, resulting in a somewhat lower confidence. This is not only a clustering problem but also a decision tree algorithm which generates a tree of incorrect accuracy with no leaf node except root nodes. The reason for this is that the attributes of the data are not suitable for the tree format and the colon cancer data need different analysis method than the decision tree algorithm. In order to obtain more accurate data afterwards it is required to analyze the data first to calibrate the appropriate data according to the format and to find the appropriate algorithm so that the desired analysis result can be obtained.

CONCLUSION

In recent years, data mining has attracted attention in order to extract information that can be practically used based on a vast amount of data held in various fields. There are various methods to analyze data but the technique and experiment results can vary greatly depending on the content of the data or characteristics of the property. In order to obtain accurate experimental results, the contents of the data should be more clearly organized and then experimented. Therefore, in this study, we analyze the data using decision tree algorithm which is one of the data mining techniques and divide the cluster of colon cancer data using clustering. Based on the change of decision tree algorithm and the accuracy of clustering based on the data.

Due to the variety of attributes, the size of the tree was not exactly accurate and the error rate was large. Therefore, in order to analyze the colon cancer data, it would be more effective to find and analyze an appropriate algorithm other than a tree analysis method. However, as the number of numerical attributes in the data increases with the information of the data itself it can be interpreted that the operation is performed more often and the examination is performed more frequently which shows different results in each region. The amount of data and the nature of the tree were not exactly the size of the tree but the clustering was divided into five clusters. However, in order to obtain more accurate results, more detailed experiments are required as well as property characteristics and size setting. In future experiments, we will look for algorithms that match the data in detail, prepare more experimental data and apply various algorithms accordingly to show the results accurately.

RECOMMENDATIONS

In the future, for more accurate experimental results it is necessary to analyze the contents of the dataset in detail to summarize the algorithms, so that, the algorithms are accurate to compare the results with other data and to analyze the common points and differences. Through comparison, it is necessary to derive more accurate and detailed results, analyze experiment results using various other techniques and derive desired results through data prediction and analysis.

ACKNOWLEDGEMENTS

This research is supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT on Development of heterogeneous big data integration and processing technology for cancer care service (2017M3C4A7083412)

REFERENCES

Davis, J., V.S. Costa, I.M. Ong, D. Page and I. Dutra, 2004. Using Bayesian Classiers to Combine Rules. In: Knowledge-Intensive, in Teractive and Efficient Relational Pattern Learning, Rickard, E. (Ed.). University of Wisconsin-Madison, Madison, Wisconsin, pp: 124-137.

Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, CA..

Hastie, T., R. Tibshirani and J. Friedman, 2001. The Elements of Statistical Learning Data Mining, Inference and Prediction. Springer, Berlin, Germany, Pages: 739.

Lamma, E., F. Riguzzi and S. Storari, 2006. Improving the K2 Algorithm using Association Rule Parameters. In: Modern Information Processing from Theory to Applications, Bouchon-Meunier, B., G. Coletti and R.R. Yager (Eds.). Elsevier, Amsterdam, Netherlands, ISBN:978-0-444-52075-3, pp: 207-217.

Moran, S., Y. He and K. Liu, 2009. An empirical framework for automatically selecting the best bayesian classifier. Proceedings of the World Congress on Engineering Vol. 1 (WCE'09), July 1-3, 2009, International Association of Engineers (IAENG), London, England, UK., ISBN:978-988-17012-5-1, pp: 1-6.

Ruiz, C., 2009. Illustration of the K2 algorithm for learning Bayes net structures. USA. http://web.cs.wpi.edu/~cs539/s07/Projects/k2_algorithm.pdf.

Witten, I.H., E. Frank and M.A. Hall, 2011. Data Mining Practical Machine Learning Tools and Techniques. 3rd Edn., Morgan Kaufmann Publishers, Burlington, Massachusetts, USA., ISBN:978-0-12-374856-0, Pages: 629.