Open vs. Close Source Decision Tree Algorithms: Comparing Performance Measures of Accuracy, Sensitivity and Specificity

Sushmita Khan
Sk03732@georgiasouthern.edu

Hayden Wimmer
hwimmer@georgiasouthern.edu

Georgia Southern University
Statesboro, GA, 30460, USA

Loreen Powell
lpowell@bloomu.edu
Bloomsburg University
Bloomsburg, PA, 17815, USA

Abstract

Data Science research is trending due the abundance of publicly available data and open source and close (proprietary) tools available. Currently, an abundant amount of research exists on various data science techniques, tools and mining of medical data and big data. However, there is little to non-existent research, which actually compares closed and open source algorithms. This research compared a closed source algorithm (Microsoft Decision Tree) with open source algorithms (CART and C4.5) performances for accuracy, sensitivity, and specificity using data form the U.S. government's Surveillance, Epidemiology, and End Results Program (SEERS). Data was downloaded, converted from raw data to structured data using a custom designed python script and transformed via the removal of missing and irrelevant data, and outliers. Predictive modeling results for accuracy, sensitivity, and specificity, indicated that closed algorithms have the best accuracy and specificity.

Keywords: Predictive Modeling, Decision Tree Algorithms, SEERS, CART, C4.5, Microsoft Decision Trees

1. INTRODUCTION

Data science is one of the most trending topics within the information and technology management field (Davenport & Patil, 2012) due to its processing and analytic ability to understand, explain, and generate insights and predictive models from various types of data (George, Osinga, Lavie, & Scott, 2016). This growing field is profoundly useful for organizations of all types, including medical organizations.

One of the most common data science usages within the medical fields is data mining. Specifically, various data mining techniques have been used throughout medical studies to aid in medical decision making, find an efficient

mechanisms, detect and diagnose medical conditions, classification of medical images and conditions as well as to predict survivability and cure rates and many other medical problems (Al-Bahrani, Agrawal, & Choudhary, 2013; Bradburn & Zeleznikow, 1994; Chun, Kim, Hahm, Park, & Chun, 2008; Detrano et al., 1989; Lu, Hales, Rew, & Keech, 2016; Nam & Shin, 2013; Shouman, Turner, & Stocker, 2011; Zheng, Yoon, & Lam, 2014)

In addition to using data science mining techniques, many research studies have taken a more technical approach by focused on developing new techniques and algorithms, as well as comparing algorithms for accuracy (Singh & Gupta, 2014). A recent study examined data mining classifiers based on a breast cancer data set and the WEKA software (Al-Hagery, 2016). They found the Bayes Net classifier model to be the most accurate model for breast cancer data sets.

Similarly, another theoretical comparative study (Singh & Gupta, 2014) explored three commonly used decision tree algorithms. Their study provided advantages and disadvantages of using the ID3, CART, and C4.5 decision tree algorithms. They also theoretically compared characteristic criteria such as splitting criteria, attribute type, pruning strategy, and outlier detection for each algorithm. Their study is limited in that it only did a theoretical comparison of open source algorithms. They did not examine commonly used close source algorithms such as Microsoft Decision Tree.

While there are numerous research studies on various data science techniques, tools and mining of medical data and big data, there is little to non-existent research, which actually compares closed and open source algorithms. This research builds upon existing research by comparing technical performances of closed source algorithm (Microsoft Decision Tree) with open source algorithm (CART and C4.5) for accuracy, sensitivity, and specificity using data from the U.S. government's Surveillance, Epidemiology, and End Results Program (SEERS). The remainder of this paper is organized as follows: background, purpose statement, methods, results and conclusion.

2. BACKGROUND

Predictive Modeling

Predictive Modeling aims to determine what is likely to happen in the future (Sharda, Delen, Turban, Aronson, & Liang, 2014). Typically, predictive modeling uses data mining techniques, which includes classification algorithms. While there are many classification algorithms, a few of the more popular are decision tree, neural networks, case-based reasoning, Bayesian classifiers, and Genetic algorithms (Sharda et al., 2014).

Data mining with advanced algorithms are very popular for the advantage of pattern discovery. Currently, data mining tools have played an essential role for predictive research within the medical field (Rani, 2014). Specifically, predictive studies have been conducted on different types of cancer, such as breast cancer, lung cancer, and colon cancer (Al-Bahrani et al., 2013; Khan, Choi, Shin, & Kim, 2008; Shen, Yang, & Shao, 2014; Xiong, Kim, Baek, Rhee, & Kim, 2005; Zheng et al., 2014). Some studies predicted the survivability of cancer patients (Al-Bahrani et al., 2013; Bellaachia & Guven, 2006; Delen, Walker, & Kadam, 2005; García-Laencina, Abreu, Abreu, & Afonoso, 2015; Khan et al., 2008) while others focused on considering the existing variables in the data set to predict the possibilities of cancer (Shen et al., 2014; Wang & Yoon, 2015; Zheng et al., 2014).

Decision Trees

A Decision Tree is a popular algorithm that graphs every possible effect to a decision. Specifically, a decision tree contains root, branch and leaf nodes. A root node is the starting pointing upon which various branches and leaf nodes occur. A branch node is a choice between various scenarios of decisions or outcomes. A leaf node is a decision (Peng, Chen, & Zhou, 2009).

Decision Trees are primarily used for prediction purposes (Mohan, 2013; Rani, 2014). Many research studies use Decision Trees because they are easy to understand (Mohan, 2013). A few widely used decision tree algorithms are CART, C4.5 and Microsoft Decision Tree.

CART

CART is a popular open source decision tree (Singh & Gupta, 2014). CART produces binary decision trees. As the term binary suggests, each branch has two nodes only. In the CART algorithm, the target variable is usually

categorical. The tree identifies in which class the output might potentially belong. The CART algorithm splits the nodes by recursive partitioning. It conducts an exhaustive search for all variables and all splitting values and then selects the optimal split The optimal split is the split that maximizes the measure of goodness over all the other splits (Sharda et al., 2014). Equation 1 is the equation for the measure of goodness of fit and equations 2 and 3 provide additional detail into the goodness of fit calculation in equation 1.

$$\emptyset(s|t) = 2P_L P_R \sum_{j=1}^{\#classes} |P(j|t_L) - P(j|t_R)|$$

Equation 1: Calculation of Goodness of Fit

Where, t_L = left child node of node t and t_R = right child node of node t

$$P_L = \frac{Number\ of\ records\ at\ t_L}{Number\ of\ records\ in\ a\ training\ set} \qquad P_R = \frac{Number\ of\ records\ at\ t_R}{Number\ of\ records\ in\ training\ set}$$

Equation 2: Additional detail for goodness of fit

$$P(j|t_L) = \frac{Number\ of\ class\ j\ records\ at\ t_L}{Number\ of\ records\ at\ t} \quad P(j|t_R) = \frac{Number\ of\ class\ j\ records\ at\ t_R}{Number\ of\ records\ at\ t}$$

Equation 3: Additional detail for goodness of fit

C4.5

C4.5 is another popular open source decision tree (Singh & Gupta, 2014). It's a decision tree algorithms and much alike CART, it recursively visits the nodes to find the optimal split However, unlike CART, C4.5 can have multiple branches and for categorical variables, it produces a separate branch for each value in the categorical attribute (Singh & Gupta, 2014). For the optimal split, C4.5 uses Information gain and Entropy (Sharda et al., 2014). The higher the information, the better split. The worst possible value for Entropy is 1. Equations 4 and 5 detail the calculation for information gain and Entropy:

$$Information\ gain = H(T) - H_s(T)$$

Equation 4: Information gain formula

Where $H_s(T) = -\sum_{i=1} H_s(T_i)$
And $Entropy = \sum_j p_j \log_2 p(j)$
(Where j is the total number of variable occurrences.)
Equation 5: Entropy formula

Microsoft Decision Tree

The Microsoft Decision Tree is another popular decision tree algorithm. Similar to CART and C4.5, however, it is proprietary/close source. The Microsoft Decision Trees algorithm uses a Bayesian approach as a tree is built by determining the correlations between the input and the targeted outcome (Microsoft, 2017).

2.3 Performance Measurements

While it is important for predicting and finding a relationship of attributes, it is also equally important to ensure the performance measures of the data mining techniques and the models used. A standard performance estimation model typical used contains three criteria: accuracy, sensitivity and specificity (Cheng, Sen, Jernigan, & Kloczkowski, 2009; Huang, McCullagh, Black, & Harper, 2004; Kantardzic, 2011; Sharda et al., 2014). Equation 5, 6, and 7 lists the formulas for each criteria.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

Equation 6: Performance estimation model formula for accuracy

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Equation 7: Performance estimation model formula for sensitivity

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

Equation 8: Performance estimation model formula for specificity

3. PURPOSE STATEMENT

The purpose of this research is to builds upon existing research by comparing technical performances of a closed source algorithm (Microsoft Decision Tree) and two open source algorithms (CART and C4.5) for accuracy, sensitivity, and specificity using data from the U.S. government's Surveillance, Epidemiology, and End Results Program (SEERS). This research will aid medical facilities and researchers in selecting the highest performing decision tree algorithm for SEER data and alike data sets.

4. METHODS

Data Source

For the purposes of this study, the data was obtained from the SEER Cancer Incidence Public-Use Database (National Cancer Institute, 2014). The SEER database is publicly available and numerous studies have been based on its data. Primarily, the database is used for analytical research in various organizations across the country. The SEER program has 9 participating registries (9 different geographic areas across the Unites States) and collects data for incidence and survival. These data sets, data dictionaries, and relevant documentation may be distributed upon request.

The SEER program is reputed for its emphasis on quality and comprehension. The data are estimated to be approximately 98% complete for each dataset. The SEER repository comprises cancer data from the years 1973 to 2013 for each registry. Amongst these years, it has data for Breast cancer, Colon and Rectum, Other Digestive, Female Genital, Lymphoma of all sites and Leukemia, Male Genital, Respiratory, Urinary and all other sites as well as the population index of each registry (National Cancer Institute, 2014). The SEER database is a popular and widely used database for acquiring cancer-related data. Several studies have used this database for the purposes of data collection (Agrawal & Choudhary, 2011; Al-Bahrani et al., 2013; Bellaachia & Guven, 2006; Delen et al., 2005; Lee, Agrawal, & Choudhary, 2013; Rathore, Tomar, & Agarwal, 2014; Wang & Yoon, 2015). As with most data, the SEER data requires formatting and preprocessing before it can be used. In order to preprocess the data, frequently a script is written and the data is exported in CSV format for further use as demonstrated in (Al-Bahrani et al., 2013).

In order to access the data an account was created and the data request application was filed. Upon being granted access, the data for Respiratory cancer was chosen for the years 2009 to 2013. The files in the record relate to the specific cancer type. The SEER Respiratory cancer incidence data consists of four datasets named:

- yr1973_2013.seer9,
- yr1992_2013.sj_la_rg_ak,
  yr2000_2013.ca_ky_lo_nj_ga and
- yr2005.lo_2nd_half.

The dataset yr1973_2013.seer9 has data for the registries Atlanta, Connecticut, Detroit. Hawaii, Iowa New Mexico, San Francisco-Oakland, Seattle, Puget, and Utah. The dataset, yr1992_2013.sj_la_rg_ak, consists data for San Jose-Monterey, Los Angeles, Rural Georgia, Alaska, YR2000_2013.CA_KY_LO_NJ_GA consists data for Grater California, Kentucky, Louisiana, New Jersey, and Greater Georgia and YR2005.LO_2ND_HALF consists data from July to December 2005 for Louisiana. Each dataset approximately 1.2 million rows of data and 72 attributes significant to respiratory cancer, which are same for all the four datasets.

Data Preparation

The data obtained from SEER was in raw unstructured format. The data had no headers, columns or rows. In other words, the data was neither readable nor useable. In order to make the data readable, a python script was written and the data was converted into the comma-separated version. The headers are typed into the python a script and the text file obtained from SEER was imported into the script. The script, along with adding the header to the data, separated the data for each attribute in the row according to the length of a character of each attribute. After each attribute in the rows, a comma is inserted. The location of the text file is provided to the python script so the data could be separated and the location where the output CSV file was to be saved was provided. After running the script, a comma-separated version of the data is available. The data is then imported into Microsoft Excel for further analysis before being imported into Microsoft SQL server management studio (SSMS) and SQL server analysis services (SSAS).

The dataset had 72 different attributes relevant to respiratory cancer over the years 1973 to 2013. The malignancy of the tumor is identified as the target variable or the dependent variable for this study. The primary ID associated with each patient is the primary key. After observing and studying the data closely, the decision is to conduct the study on data for the years 2009 to 2013. Hence there is a focus on the most recent 5 years. Seeing how cancer is an ever progressing disease with the factors circulating and influencing continuously evolving, the idea was to address data as recent as possible to understand the malignancy consequently cancers effect with respect to the data for the attributes in the recent times as opposed to 20 to 30 years back. Figure 1, located in the Appendix shows an illustration of the extraction process of data from SEER.

Twenty attributes were removed from the data set which has no data whatsoever for the years 2009 to 2013. This 20 attribute involves different measures of the tumor size, tumor markers, recording of lymph nodes with a conjunction of the surgery performed, surgery, radiation, few stages of AJCC edition and SEER Summary stages. In addition to these data, some other constructs were removed which were not deemed significant or were considered repetitive. However, even after removing these attributes, the dataset still has missing data for fourteen attributes. The data for these fourteen attributes are missing only for the year 2009. Seeing how the majority of the dataset has the values for Collaborative stage site-specific factor and Derived AJCC (American Joint committee on cancer) these attributes are included in the data set.

The data set is cleaned of further irrelevant attributes. This is done prevent the effect on the result of the study. Attributes with data for brain cancer, liver cancer, breast cancer and bone cancer were removed. They are not relevant to the lung cancer study, have spurious data and in order to make sure the result is not influenced by any irrelevant data, these data were removed.
Attributes like CS3SITE, CS4SITE, CS5SITE and CS6SITE has no data for the selected years, except for 3672 rows out of the total of 93,458 rows approximately. With a 0.04% of available data in each four columns, these attributes were removed. Regardless of removing these attribute, several other attributes for example DAJCCT, DAJCCZ, CSMETSDX_LUNG still has missing values. However, they were not removed, seeing their relevance to the data set. Additionally, these variables were missing data for 18,000 rows each approximately.

Model Planning and Building

For the purposes of this study, the applications SSMS and SQL SSAS were selected to analyze the data set using a closed source decision tree algorithm, Microsoft Decision Tree. Microsoft SSAS provides an array of data mining techniques including a Decision tree, Artificial Neural network, and Naïve Bayes algorithms (Microsoft, 2016). For the purposes of this study, the decision tree algorithm was selected as the data mining technique. As suggested by Microsoft Developer Network (MSDN), 30% of the data was reserved for a test set. Additionally, the input variables in CART and C4.5 were the same as input to SSAS.

5. RESULTS

Closed Sourced Algorithm (Microsoft Decision Tree)

The Microsoft Decision Tree has 13 levels and 51 rules. The levels are a combination of the variables and the rules generated to find the relationship with the output variable. The training set comprises of 63894 cases. Of this number 48573 are classified as 1 or as the tumor being malignant and 15321 are classified as 0 that is nonmalignant. A total of 76.01% of the cases are classified as malignant. Figure 2, in the Appendix, shows the resulting decision tree from the Microsoft Decision Tree algorithm.

When running the Microsoft Decision Tree, a confusion matrix, lift chart and a dependency network was generated by SSAS. The classification matrix was used to calculate the accuracy of the overall prediction. It was found, using the formulas of accuracy that, this algorithm has an accuracy of 77%. The sensitivity is 3.62% and the specificity is 99.8%. 70933 instances were classified correctly and 22545 instances were classified incorrectly. The lift chart generated reflects upon the percentage of the correctly classified classes as a comparison with the ideal chart. Additionally, using the mining model prediction, the chart generated showed the classes which were classified correctly and the classes that were classified was classified incorrectly. In the lift chart, the top line is the ideal scenario generated by SSAS and the bottom line is an actual prediction.

The lift chart is the demonstration of the comparison of prediction of the models against mining results generated by random guesses using the available data. From a looking at the lift chart, it can be seen that the mining results generated by the Microsoft Decision Tree model have a lower gradient than the mining results of the random guess mode. The lift for the model was computed using the confusion matrix. Results found that the model has a lift of 36%. However, the life for the results generated randomly is 50%. The following is the formula for computing the lift of the model:

$$Lift\ for\ Model = \frac{\frac{True\ Positive}{True\ Positive + False\ Negative}}{\frac{True\ Positive + False\ Positive}{Total\ number\ of\ Instances}}$$

Equation 6: Lift Calculation

Open Source Algorithms (C4.5 and CART)

C4.5 Algorithm

C4.5 generated a confusion matrix for each of the algorithm. The performance of these algorithms was calculated using the same measure as the Microsoft Decision Tree. Accuracy, sensitivity, and specificity were used. In C4.5, specificity was called Precision and Sensitivity was called Recall. The C4.5 algorithm classified 71,274 instances correctly and 22204 instances were classified incorrectly. The tree has 101 leaves and the size of the tree is 201. Given the number of inputs and total rows of the data set, a tree as large as this was anticipated. A classification matrix was generated, along with the time required to analyze the data. A classification matrix is also referred to as a confusion matrix.

The measures accuracy, specificity, and sensitivity were calculated to evaluate the performance of the C4.5 algorithm. The values for true positive, true negative, false positive and false negative are determined from the confusion matrix. Upon calculation of accuracy, specificity, and sensitivity, this model has an accuracy 76.23%, sensitivity is 64.4% and specificity is 98.8%. The C4.5 model required 4.34 seconds to build the decision tree. The root mean square error has a value of 0.4206 and the mean absolute error has a value of 0.3524. The root mean square error is greater than the mean absolute error as well as the fact these values are relatively small. The relative absolute error is 94.3% and the root relative squared error is 97.3%.

CART

The CART models were evaluated by the measures accuracy, specificity, and sensitivity. The values generated from the confusion matrix was used to perform the calculations. The specificity for this model was 65.6% and the sensitivity of this model was 9.2%. The accuracy of this model is 76.2%. Of all the 93478 attributes, 71268 were classified correctly and the remaining 22210 was classified incorrectly. The mean absolute value for this model is smaller than the root means square error. The mean absolute error has a value of 0.3527 and the root mean square error has a value of 0.4199. A relative absolute error has a value of 94.4% and root relative squared error has a value of 97.2%. The values for mean absolute error and root mean square error both has relatively small values for this model.

Table 1 show a comparison of all three decision tree algorithms. As illustrated, Microsoft Decision Tree preformed the best for accuracy and specificity.

Table 1: Performance Measure for Open vs. Close Decision Tree Algorithms

	Accuracy	Sensitivity	Specificity
Microsoft Decision Tree	77%	4%	100%
C4.5	76%	64%	99%
CART	76%	9%	66%

6. CONCLUSION AND DISCUSSION

This work examines the performance of open versus closed source decision tree algorithms, namely Microsoft Decision Tree with CART and C4.5. Accuracy is computed for all three models. The difference in accuracy for these models are nominal; however, in a medical scenario a nominal disparity can be the difference between life and death. Therefore, if the models are to be ranked in ascending order of accuracy, Microsoft Decision Tree has the best accuracy, followed by C4.5 and then the CART algorithm. Given the percentages for accuracy, sensitivity, and specificity, it can be inferred that the three algorithms perform with a similar level of accuracy. However, Microsoft Decision Tree, although it has a very low sensitivity and extremely high specificity, has the highest accuracy. Thus, it can be said that Microsoft Decision Tree, in terms of accuracy, is the better performing algorithm. Given the high specificity of the Microsoft Decision Tree, it predicts all output correctly. This holds true for C4.5 as well, but because the specificity percentage is low, the prediction power of C4.5 is weaker than that of the Microsoft Decision Tree. After comparison of accuracy sensitivity and specificity, Microsoft Decision Tree, although by 0.8% and 0.77%, better predicted the outcomes. The Microsoft Decision Tree also outperforms CART and C4.5 on specificity percent and sensitivity. Based on the results of this research, medical facilities and researchers should consider closed source options in data science and predictive analytics as the Microsoft Decision Tree was the highest performing decision tree algorithm on SEER data. Future research intends to address the Microsoft Decision Tree on more diverse datasets and compare performance in terms of efficiency of

data processing in preparation for larger datasets. In conclusion, this work shows that, while most research focuses on open source algorithms, researchers should consider closed source as it may offer additional improvements or add another dimension to aid clinicians in evidence based medicine and clinical decision support.

Compliance with Ethical Standards:

Authors 1, 2, and 3 declare that he/she has no conflict of interest. This article does not contain any studies with human participants or animals performed by any of the authors.

7. REFERENCES

Agrawal, A., & Choudhary, A. (2011). *Identifying HotSpots in lung cancer data using association rule mining.* Paper presented at the Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on.

Al-Bahrani, R., Agrawal, A., & Choudhary, A. (2013). *Colon cancer survival prediction using ensemble data mining on SEER data.* Paper presented at the Big Data, 2013 IEEE International Conference on.

Al-Hagery, M. A. H. (2016). Classifiers' Accuracy Based on Breast Cancer Medical Data and Data Mining Techniques. *INTERNATIONAL JOURNAL OF ADVANCED BIOTECHNOLOGY AND RESEARCH, 7*(2), 760-772.

Bellaachia, A., & Guven, E. (2006). 1 Predicting Breast Cancer Survivability Using Data Mining Techniques.

Bradburn, C., & Zeleznikow, J. (1994). The application of case-based reasoning to the tasks of health care planning. *Topics in case-based reasoning*, 365-378.

Cheng, H., Sen, T. Z., Jernigan, R. L., & Kloczkowski, A. (2009). Data Mining for Protein Secondary Structure Prediction *Data Mining in Crystallography* (pp. 59-87): Springer.

Chun, S. C., Kim, J., Hahm, K. B., Park, Y. J., & Chun, S. H. (2008). Data mining technique for medical informatics: detecting gastric cancer using case-based reasoning and single nucleotide polymorphisms. *Expert Systems, 25*(2), 163-172.

Davenport, T. H., & Patil, D. (2012). Data scientist. *Harvard Business Review, 90*, 70-76.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine, 34*(2), 113-127.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., . . . Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology, 64*(5), 304-310.

García-Laencina, P. J., Abreu, P. H., Abreu, M. H., & Afonoso, N. (2015). Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Computers in biology and medicine, 59*, 125-133.

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods for management research. *Academy of Management Journal, 59*(5), 1493-1507.

Huang, Y., McCullagh, P., Black, N., & Harper, R. (2004). Evaluation of outcome prediction for a clinical diabetes database *Knowledge Exploration in Life Science Informatics* (pp. 181-190): Springer.

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*: John Wiley & Sons.

Khan, M. U., Choi, J. P., Shin, H., & Kim, M. (2008). *Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare.* Paper presented at the Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE.

Lee, K., Agrawal, A., & Choudhary, A. (2013). *Real-time disease surveillance using twitter data: demonstration on flu and cancer.* Paper presented at the Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.

Lu, J., Hales, A., Rew, D., & Keech, M. (2016). *Timeline and episode-structured clinical data: Pre-processing for Data Mining and analytics.*

Paper presented at the Data Engineering Workshops (ICDEW), 2016 IEEE 32nd International Conference on.

Microsoft. (2016). Data Mining (SSAS). Retrieved from https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/data-mining-ssas

Microsoft. (2017). Microsoft Decision Tree Algorithm Technical Reference. Retrieved from https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/microsoft-decision-tree-algorithm-technical-reference

Mohan, V. (2013). Decision Trees: A comparison of various algorithms for building Decision Trees.

Nam, Y., & Shin, H. (2013). *A hybrid cancer prognosis system based on semi-supervised learning and decision trees.* Paper presented at the International Conference on Neural Information Processing.

National Cancer Institute. (2014). *SEER Research Data*.

Peng, W., Chen, J., & Zhou, H. (2009). An implementation of ID3-decision tree learning algorithm. *From web. arch. usyd. edu. au/wpeng/DecisionTree2. pdf Retrieved date: May, 13*.

Rani, P. S. (2014). A Study on Data Mining Classification Algorithms For Medical Data. *International Journal of Advanced Research in Computer Science, 5*(2).

Rathore, N., Tomar, D., & Agarwal, S. (2014). *Predicting the survivability of breast cancer patients using ensemble approach.* Paper presented at the Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on.

Sharda, R., Delen, D., Turban, E., Aronson, J., & Liang, T. P. (2014). *Businesss Intelligence and Analytics: Systems for Decision Support-(Required)*: Prentice Hall.

Shen, R., Yang, Y., & Shao, F. (2014). *Intelligent breast cancer prediction model using data mining techniques.* Paper presented at the Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2014 Sixth International Conference on.

Shouman, M., Turner, T., & Stocker, R. (2011). *Using decision tree for diagnosing heart disease patients.* Paper presented at the Proceedings of the Ninth Australasian Data Mining Conference-Volume 121.

Singh, S., & Gupta, P. (2014). Comparative study ID3, cart and C4. 5 decision tree algorithm: a survey.

Wang, H., & Yoon, S. W. (2015). *Breast Cancer Prediction Using Data Mining Method.* Paper presented at the IIE Annual Conference. Proceedings.

Xiong, X., Kim, Y., Baek, Y., Rhee, D. W., & Kim, S.-H. (2005). *Analysis of breast cancer using data mining & statistical techniques.* Paper presented at the Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks. SNPD/SAWN 2005. Sixth International Conference on.

Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications, 41*(4), 1476-1482.

APPENDIX

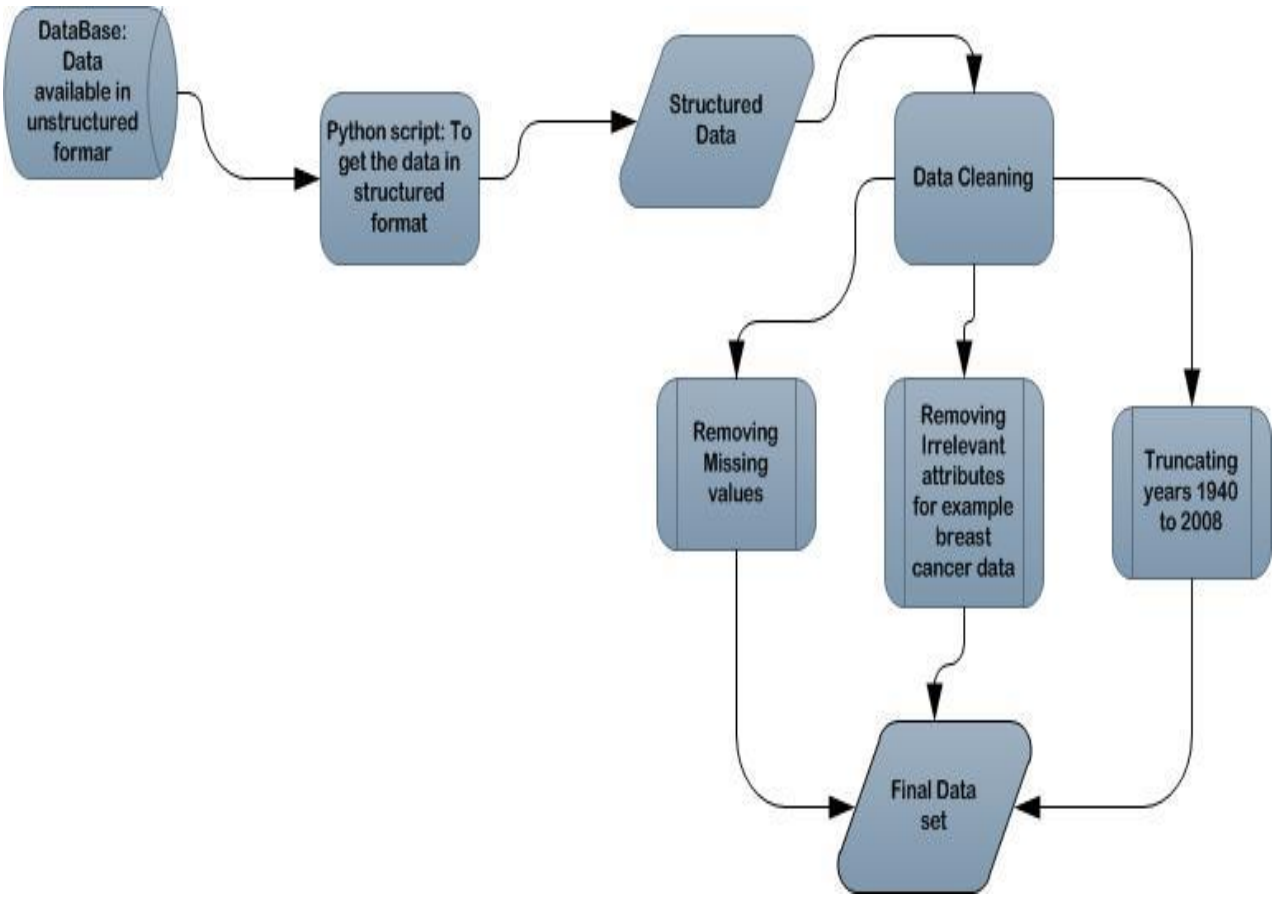

Figure 1: Extracting Data from SEER

ISSN: 2167-1508
v10 n4504

Figure 2: Microsoft Decision Tree