Comparison of Performance of Decision Tree Algorithms and Random Forest: An Application on OECD Countries Health Expenditures

Songul Cinaroglu
Hacettepe University
Faculty of Economics and Administrative Sciences
Department of Health Care Management Beytepe Ankara

ABSTRACT

Decision trees and Random Forest are most popular methods of machine learning techniques. C4.5 which is an extension version of ID.3 algorithm and CART are one of these most commonly use algorithms to generate decision trees. Random Forest which constructs a lot of number of trees is one of another useful technique for solving both classification and regression problems. This study compares classification performances of different decision trees (C4.5, CART) and Random Forest which was generated using 50 trees. Data came from OECD countries health expenditures for the year 2011. AUC and ROC curve graph was used for performance comparison. Experimental results show that Random Forest outperformed in classification accuracy [AUC=0.98] in comparison with CART (0.95) and C4.5 (0.90) respectively. Future studies more focus on performance comparisons of different machine learning techniques using several datasets and different hyperparameter optimization techniques.

General Terms

Pattern Recognition, Machine Learning, Decision Trees.

Keywords

Pattern Recognition, Machine Learning, Decision Trees, Health Expenditures.

1. INTRODUCTION

Machine learning techniques are used for data analysis and pattern discovery. Thus play a major role in the development of data mining applications [1]. These techniques aim to develop algorithms in the analysis of large complex datasets [2] common strategy of these techniques is to discover a pattern in a training dataset [1]. Classification is an important part of machine learning and data mining applications; it defines groups within population. There are many different methods to compare results and to determine the best classification [3]. Rule induction which is one of machine learning techniques creates decision trees or a set of classification rules from training examples with a known classification [1].

A tree is a data representation model and originated from a graph theory [4]. Graph theory is a mathematic based theory and it is applied to solve any practical problems [5]. Trees used in data structures, data bases, computer algorithms, machine learning and data mining. This method produces a nonparametric classification and prediction model. Decision trees are nonlinear supervised learning models and these models use the concept of trees [4]. They work with nominal outcomes that have more than two possible results and with ordinal outcome variables [3]. These models organized in the form of a rooted tree with two types of nodes. These nodes called decision and class nodes [4]. Number of algorithms have been proposed for decision tree construction. One of these algorithms is ID.3 (Interactive Dichotomizer 3). This algorithm is an effective and popular method for finding decision tree rules [6]. Information gain is exactly the metrics for selecting the best attribute in each step of the growth tree in ID3 algorithm [7]. C4.5 is another algorithm using in decision trees. This is an extension version of ID.3 [4]. Both of them are most popular in the machine learning community [8]. The difference between these two algorithms is that ID.3 uses binary splits, C4.5 uses multi-way splits. CART (Classification and Regression Tree) is another induction algorithm using in decision tree models. It produces a regression tree when outcome is continuous and classification tree when the outcome is categorical. It is well suited to the generation of clinical decision rules [4]. Table 1 shows comparison of common decision tree induction algorithms [4]. Whereas CART and C4.5 produce general trees, ID.3 produce binary decision trees. Split criteria for CART and ID.3 is 2-way, it is multiway for C4.5. Additionally, ID.3 and C4.5 uses entropy for the induction measure but CART uses Gini coefficient for that. The oldest one is CART algorithm developed in 1984, ID.3 is another algorithm developed in 1986, finally C4.5 developed in 1993.

Table 1. Comparison of common decision tree induction algorithms

Algorithm	Attribute	Split	Measure	Year
CART	General	2-way	Gini	1984
ID.3	Binary	2-way	Entropy	1986
C4.5	General	Multi-way	Entropy	1993

Random Forest is another machine learning method, uses CART algorithm [9]. This technique is useful for classification, regression and other tasks, it constructs many decision trees. Random Forest will be used to classify a new instance by the majority vote [10]. Determining optimal number of trees in a Random Forest is still an open question [11]. Some studies which compare Random Forest performance results using different kinds of datasets, show that as the number of trees grows, it does not always mean that the performance of the forest is better than forests which have fever trees. In other words, large number of trees only increases computational costs but not performance results [12]. There are number of performance measures for making a comparison between decision tree algorithms. Area under ROC curve (AUC) is one of these performance measures [13] and accuracy of various decision tree classifiers are compared using ROC curve which is one of performance measure for classification accuracy. This method is useful for organizing classifiers and visualizing their performance. ROC graphs are commonly used in medical decision making, in recent years they have been used in machine learning and data mining research [14]. In medical sciences a person is assessed as diseased (positive) and healthy (negative) depending on whether the corresponding marker value is greater than or less than or equal to a given threshold value. The theoretical ROC curve is a plot of q=sensitivity versus p=1-specificity for all possible threshold values ROC curve areas are typically between 0.5 and 1.0. If this value equal to 1.0 this means that this test is 100% accurate because both the sensitivity and specificity are 1.0 so there are no false positives and no false negatives [15]. In this study decision tree algorithms and Random Forest classification performance results compared with using ROC curve as a performance measure. OECD countries health expenditure dataset was used for classification. Health care expenditures have steadily increase in OECD countries. Because of that increase in these expenditures have attracted attention of politicians and health economists [16]. Totally 34 OECD countries classified according to their total health expenditure for the year 2011. Indicators related with health status, health care resources, health care utilization, nonmedical determinants of health and economic references are used for classification.

2. METARIALS & METHODS

2.1 Aim

The aim of this study is to make a comparison between machine learning techniques classification performances using AUC. For this aim C4.5, CART and Random Forest performance results compared which are most popular decision tree algorithms in machine learning on OECD countries health expenditure dataset.

2.2 Data & Analysis

Data came from OECDStatExtracts [17] to predict OECD countries total health expenditures for the year 2011. Table 2 shows name, group, explanation, measurement unit, type of each variable. Total health expenditure which is a predictor variable in this study measures; the consumption of health services and goods, outpatient care, hospital care, long-term care, pharmaceuticals, other medical goods, prevention, public health services and administration. Independent variables are; life expectancy at birth, perceived health status, number of physicians, number of hospitals, hospital aggregates, immunization, alcohol consumption and GDP per capita.

In this study analysis performed in Orange data mining program. Decision trees generated using C4.5 and CART algorithms. After decision tree generation, Random Forest was performed, number of trees was determined 50. k=5-fold cross validation was used during decision tree generation process. For comparing results in terms of classification accuracy of decision trees and Random Forest, AUC performance results and graphical representation of ROC Curve was used.

3. FINDINGS

3.1 Descriptive Statistics

Table 3 shows descriptive statistics of study variables used for predicting health expenditures in OECD countries for the year 2011. In 2011, average spending on health goods and services per person was 3394,44 (±1597,56) $. According to the mean values of predictive variables; life expectancy at birth was 80,15 (±2,46) (year); perceived health status was 66,94 (11,51) (%); total number of physicians was 113310,32 (±143735,72); number of hospitals was 1280,65 (±1862,56); hospital aggregates was 4664635,68 (±4536116,94); immunization was 96,06 (±3,20) (%); alcohol consumption per person was 9,35 (2,11) finally GDP per capita was 35436,06 (±13802,03) ($).

Table 2. Explanation of study variables

Variable Name	Group of Variable	Explanation	Measurement Unit	Type of Variable	Year
Health Expenditures	Health Expenditure and Financing	Per capita US $ PPP	Purchasing Power Parity ($)	Numeric/ Continuous	2011
Life Expectancy at Birth	Health Status	Life expectancy at Birth in Total Population	Year	Numeric/ Continuous	2011
Perceived Health Status	Health Status	Good/Very Good Health (Total Aged 15+)	Percent (%)	Numeric/ Continuous	2011
Number of Physicians	Health Care Resources	Professionally Active Physicians (Number of Persons-Head Counts)	Number	Numeric/ Discrete	2011
Number of Hospitals	Health Care Resources	Total Number of Hospitals	Number	Numeric/ Discrete	2011
Hospital Aggregates	Health Care Utilization	Inpatient Care Discharges (All Hospitals)	Number	Numeric/ Continuous	2011
Immunization	Health Care Utilization	% of Children Immunized (Diphtheria, Tetanus, Pertussis)	Percent (%)	Numeric/ Continuous	2011
Alcohol Consumption	Non Medical Determinants of Health	Liters per capita (+15)	Liter	Numeric/ Continuous	2011
GDP per capita	Economic References	GDP per capita [US $ PPP]	Purchasing Power Parity ($)	Numeric/ Continuous	2011

3.2 Correlations Matrix for Predictor Variables

In this study as a part of preliminary analysis process, Spearman correlation coefficient was used to detect multicollinearity problem among predictive variables (see Table 4). Multicollinearity refers to the linear relationship among two or more variables. It is an important problem and may cause serious difficulty with the reliability of the estimates of the model parameters [18]. Before the analysis relationships between independent variables analyzed with Spearman correlation coefficient. Spearman correlations is one of multivariate non-parametric tests [19]. Table 4 shows Spearman correlations matrix of independent variables. The magnitude of 0.70 and higher indicate high correlations. All correlations are lower in this table, by means of that there is no multicollinearity problem was detected between independent variables. This shows that all independent variables are appropriate to predict health expenditures of OECD countries.

Table 3. Descriptive statistics of study variables

	Variables	N	Min.	Max.	Mean	SD.
Independent Variable	Health Expenditures	34	937	8483	3394,44	1597,56
Predictive Variables	Life Expectancy at Birth	34	74	83	80,15	2,46
	Perceived Health Status	34	34	88	66,94	11,51
	Number of Physicians	34	1121	809492	113310,32	143735,72
	Number of Hospitals	34	8	8605	1280,65	1862,56
	Hospital Aggregates	34	78704	19868738	4664635,68	4536116,94
	Immunization	34	83	99	96,06	3,20
	Alcohol Consumption	34	2	12	9,35	2,11
	GDP per capita	34	16984	88781	35436,06	13802,03

Table 4. Correlation matrix

Variables	r_s	Life Expectancy at Birth	Number of Physicians	Number of Hospitals	Hospital Aggregates	Alcohol Consumption	GDP per capita	Perceived Health Status	Immunization
Life Expectancy at Birth	r_s	1							
Number of Physicians	r_s	-0,14	1						
Number of Hospitals	r_s	0,12	0,50**	1					
Hospital Aggregates	r_s	0,15	0,45**	0,49**	1				
Alcohol Consumption	r_s	-0,17	0,22	-0,27	-0,18	1			
GDP per capita	r_s	0,52**	-0,09	-0,01	-0,15	0,02	1		
Perceived Health Status	r_s	0,58**	-0,32	-0,07	-0,08	-0,19	0,61*	1	
Immunization	r_s	-0,11	-0,08	0,11	0,12	-0,05	-0,30	-0,40*	1

r_s : Spearman Correlation Coefficient

*p<0.05
**p<0.01

3.3 Categories for Predictive Variable

Decision trees are used to predict the classes of a categorical dependent variable. For this necessity health expenditure which is a dependent variable of this study categorized into two groups. Table 5 shows two categories of dependent variable. Mean value was used for categorization of health expenditure variable. OECD countries mean value of total health expenditure under 3394,44 $ for the year 2011 was coded 1, others coded 2. According to this classification, health expenditure of countries in the second cluster higher than in the first cluster.

Table 5. Categories of total health expenditure of OECD countries

Category	Total Health Expenditure Per capita US $ PPP (2011)	Number (Number of Countries)	Percent (%)
1	<3394,44 $	17	50%
2	≥3394,44 $	17	50%

3.4 Comparison of Different Decision Tree Algorithms Performance Results

Table 6 shows performance results of decision tree algorithms and Random Forest (number of trees=50). In this study AUC was used as performance measurement criteria, when AUC=1 this implies a perfect forecast (Marzban 2004). According to these results, Random Forest which was generated using 50 trees has high AUC performance result (AUC=0.98). In addition to that decision tree which was generated using CART algorithm have higher AUC performance results than C4.5.

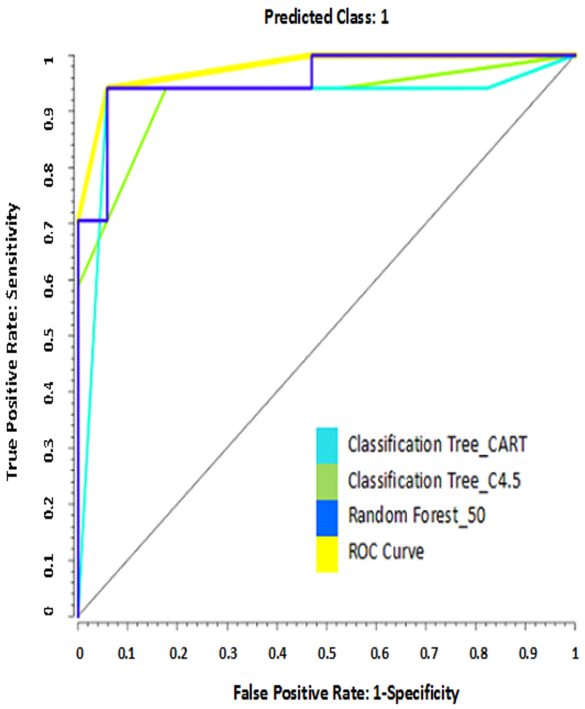
Table 6. Performance results of different decision tree algorithms and Random Forest

Decision Tree Algorithms	AUC (Area Under the ROC Curve) Performance Results
Decision Tree_C4.5	0.90
Decision Tree_CART	0.95
Random Forest [Number of Trees 50]	0.98

Graph 1 shows graphical representation of ROC curve performance results. It can be seen in this graph that Random Forest [50] (represented with dark blue) much closer to the ROC curve (represented with yellow). After that decision tree generated by CART algorithm (represented with light blue) close to ROC curve, decision tree which is generated by using C4.5 (represented with green) is the last one.

4. CONCLUSION

Classification methods was used to classify the behavior of new examples [20]. These methods are used in machine learning, data mining and multivariate statistics. What is more there is still a debate about what is the difference between data mining, machine learning and statistics. According to these debates, it is known that machine learning techniques are an important step for developing data mining techniques. If a model become successful in the development process of machine learning techniques, this points out that this model have a potential to perform well in large datasets. On the other hand, the difference between machine learning, data mining and statistics is that, statistics is far more than just hypothesis testing but many machine learning and data mining techniques do not involve any searching at all [21].



Graph 1: ROC curve

Decision trees are useful performance assessment tools for exploring the performance of different classifiers [21]. They discover rules and relationships between variables [22]. It is possible to predict the class of an example based on the values of its predictor variables [23]. Some popular algorithms used for the classification trees are CART, ID.3 and C4.5 [22]. Another popular machine learning and data mining method used for both classification and regression is Random Forest. This method composes of number of trees however determining optimal number of trees in the forest is still remain a question for Random Forest [10].

There are number of performance measures for assessing performance of decision trees and Random Forest. Area Under ROC Curve is one of these performance measures. ROC curve is a graphical plot that illustrates the performance of a binary classifier system. This is a useful performance assessment tool to explore performance of different classifiers [21]. In this study classification accuracy of decision trees and Random Forest compared with using OECD countries health expenditure data, mean value of OECD countries total health expenditures for the year 2011 was used as a cut-off point and generate a binary predictive variable. Number of independent variables related with health expenditure are used for prediction. These are; life expectancy at birth, number of physicians, number of hospitals, hospital aggregates, alcohol consumption, GDP per capita, perceived health status and immunization. As a result of the study Random Forest yields better classification accuracy Random Forest [Number of trees=50] (AUC=0.98) compared with decision trees which are generated using C4.5 (AUC=0.90) and CART (AUC=0.95) algorithms. The results obtained shows that, Random Forest has higher classification performance results than decision trees.

Despite literature suggests that large number of trees in a forest only increases computational costs and has no significant performance gain [11] it is advisable for future studies to observe whether Random Forest classification accuracy changes according to increasing number of trees in the forest. Using several datasets and looking at different kinds of performance results in addition to ROC Curve are additional recommendations for further studies to help our understanding of comparison of different machine learning techniques classification performances.

5. REFERENCES

[1] Bose, I, Mahapatra, R.K. “Business data mining - a machine learning perspective”, Information & Management, 2001, 39, 211-225.

[2] Libbrecht M.W. Noble W.S. “Machine learning applications in genetics and genomics”, Nature Reviews, 2015, 16, 321-322.

[3] Das R. “A comparison of multiple classification methods for diagnosis of Parkinson disease”, Expert Systems with Applications, 2010, 37, 1568-1572.

[4] Chattamvelli R. “Data Mining Methods, Alpha Science International”, Oxford, UK. 2009.

[5] Hammond D.K. Vandergheynst P. Gribonval R. “Wavelets on graphs via spectral graph theory”, Applied and Computational Harmonic Analysis, 2011, 30(2), 129-150.

[6] Baldwin J.F. Lawry J. Martin T.P. “A Mass assignment based ID3 algorithm for decision tree induction”, International Journal of Intelligent Systems, 1997, vol.12, 523-552.

[7] Jin C. De-lin L. Fen-Xiang M. “An improved ID3 decision tree algorithm”, Proceeding of 2009 4th International Conference on Computer Science & Education, 2009, 127-130.

[8] Salzberg S.L. “C4.5: Programs for machine learning” by Quinlan J.R. Morgan Kaufmann Publishers, Inc., 1993

[9] Gislason P.O. Benediktsson J.A. Sveinsson J.R. “Random forests for land cover classification”, Pattern Recognition Letters, 2004, 27(4), 294-300.

[10] Oshiro T.M. Perez P.S. Baranauskas J.A. “How many trees in a random forest? Machine learning and data mining in pattern recognition”, Lecture Notes in Computer Science, 2012, vol.7376, 154-168.

[11] Liaw A. Wiener M. “Classification and regression by randomforest”, R News, 2002, vol.2/3, 18-22.

[12] Latinne P. Debeir O. Decaestecker C. “Limiting the number of trees in random forests”, Multiple Classifier Systems, Lecture Notes in Computer Science, Springer. 2001.

[13] Bradley A. “The use of the area under the ROC curve in the evaluatıon of machıne learning algorithms”, Pattern Recognition, 1997, 30(7), 1145-1159.

[14] Fawcett T. “An introduction to ROC analysis”, Pattern Recognition Letters, 2006, 27(8), 861-874.

[15] Fraggi D. Reiser B. “Estimation of the area under the ROC curve”, Statistics in Medicine, 2002, 21, 3093-3106.

[16] Potrafke N. “The growth of public health expenditures in OECD countries: Do government ideology and electoral motives matter?”, Journal of Health Economics, 2010, 29, 797-810.

[17] OECD StatExtracts, http://stats.oecd.org/, Accessed on: 25.01.2016.

[18] Alin A. “Multicollinearity”, Computational Statistics, 2010, 2(3), 370-374.

[19] Oja H. Randles R.H. “Multivariate nonparametric tests”, Statistica Science, 2004, 19(4), 598-605.

[20] Dietterich T.G. “An Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization”, Machine Learning, 40, 139-157, 2000.

[21] Witten I.H. Frank E. Data Mining Practical Machine Learning Tools and Techniques, Elsevier, Third Edition, Morgan Kaufmann Publishers. 2005.

[22] Sohn S.Y. Moon T.H. “Decision Tree based on data envelopment analysis for effective technology commercialization”, Expert Systems with Applications, 2004, 26, 279-284.

[23] Rotim S.T. Dobsa J. Krakar Z. “Using decision trees for identification of most relevant indicators for effective ICT Utilization”, Bulgarian Academy of Sciences, Cybernetics and Information Technologies, 2013, 13(1), 83-94.