International Journal of Advanced Smart Convergence Vol.5 No.1 30-33 (2016)
http://dx.doi.org/10.7236/IJASC.2016.5.1.30

IJASC 16-1-5

Performance Comparison of Decision Trees of J48 and Reduced-Error Pruning

Hoon Jin*, Yong Gyu Jung**,+
**Dept. of Computer Engineering, Sungkyunkwan University, Korea*
bioagent@gmail.com

***Department of Medical IT Marketing, Eulji University, Korea*
ygjung@eulji.ac.kr

Abstract

With the advent of big data, data mining is more increasingly utilized in various decision-making fields by extracting hidden and meaningful information from large amounts of data. Even as exponential increase of the request of unrevealing the hidden meaning behind data, it becomes more and more important to decide to select which data mining algorithm and how to use it. There are several mainly used data mining algorithms in biology and clinics highlighted; Logistic regression, Neural networks, Supportvector machine, and variety of statistical techniques. In this paper it is attempted to compare the classification performance of an exemplary algorithm J48 and REPTree of ML algorithms. It is confirmed that more accurate classification algorithm is provided by the performance comparison results. More accurate prediction is possible with the algorithm for the goal of experiment. Based on this, it is expected to be relatively difficult visually detailed classification and distinction.

Keywords: *Data Mining, Weka, Maching Learning, Classifiacation, J48, REPTree*

1. INTRODUCTION

Data mining reveals the recent prominence of Machine Learning. It means to acquire the skills and knowledge that can be acquired information efficiently. Machine learning is used to predict the outcome of new information. It may be due to them from the new data. Data mining is defined as "To create a new and useful knowledge from large amounts of data", as same meaning of data processing, data summarization, machine learning, pattern recognition, visualization, statistics, knowledge extraction technique, etc requires the skills of a variety of fields. Machine Learning provides a methodology to extract the information from the source data in the database as the primary technical-based data mining. For example, if a model consisting of a parameter, on the basis of past experience or training data is referred to as a computer program or a training learning act to optimize the parameters of the model. The learned model can predict the results from the new data have never met in the learning process. In this paper, we compare the performance of the algorithm proceeds separation is made based on the ML. Machine learning is used to predict the

Manuscript Received: Nov. 17, 2015/ Revised: Dec. 5, 2015/ Accepted: Jan. 23, 2016
Corresponding Author: ygjung@eulji.ac.kr
Tel: +82-31-740-7178, Fax: +82-31-740-7198
Department of Medical IT Marketing, Korea, Eulji University

outcome of new information. It may be due to them from the new data. Undergo a process of using a known supervised algorithm and to ensure that any algorithm J48 compared to the performance of this REPTree provide more accurate classification results in the goal of this study is to find a more accurate prediction algorithms. Based on this, it is expected to be relatively difficult visually detailed classification and distinction.

2. RELATED RESEARCH

ML algorithm can consider dividing into unsupervised, supervised algorithm.. ML algorithm is unsupervised classification for each class based on the similarity of the given data, there is a K-Means, DBSCAN, AutoClass, Expectation Maximization, etc. Supervised ML algorithm is an algorithm that classifies the test data based on it is not known after training the model using the known data, there is J48, REPTree, NaiveBayes, BayesNet, etc. Decision trees represent a supervised approach to classification. A decision tree is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes. The ordinary tree consists of one root, branches, nodes (places where branches are divided) and leaves. In the same way the decision tree consists of nodes which stand for circles, the branches stand for segments connecting the nodes. A decision tree is usually drawn from left to right or beginning from the root downwards, so it is easier to draw it. The first node is a root. The end of the chain "root- branch - node-...- node" is called "leaf". From each internal node (i.e. not a leaf) may grow out two or more branches. Each node corresponds with a certain characteristic and the branches correspond with a range of values. These ranges of values must give a partition of the set of values of the given characteristic.

2.1 J48

J48 is a class algorithm for generating a C4.5 decision tree. C4.5 algorithm was proposed in 1993 by Quinlan using the concept of entropy as well as the Information Gain determines the separation criteria. Entropy means a level of congestion of a given data set, Information Gain means that better identifies the contents of data due by selecting any property. J48 is a tree based learning approach. It is developed by Ross Quinlan which is based on iterative dichtomiser (ID3) algorithm. J48 uses divide-and-conquer algorithm to split a root node into a subset of two partitions till leaf node (target node) occur in tree. Given a set T of total instances the following steps are used to construct the tree structure.

Step 1: If all the instances in T belong to the same group class or T is having fewer instances, than the tree is leaf labeled with the most frequent class in T.

Step 2: If step 1 does not occur then select a test based on a single attribute with at least two or greater possible outcomes. Then consider this test as a root node of the tree with one branch of each outcome of the test, partition T into corresponding T1, T2, T3........, according to the result for each respective cases, and the same may be applied in recursive way to each sub node.

Step 3: Information gain and default gain ratio are ranked using two heuristic criteria by algorithm J48.

2.2 REPTree

One of the questions that arises in a decision tree algorithm is the optimal size of the final tree. A tree that is too large risks overfitting the training data and poorly generalizing to new samples. A small tree might not capture important structural information about the sample space. However, it is hard to tell when a tree algorithm should stop because it is impossible to tell if the addition of a single extra node will dramatically decrease error. This problem is known as the horizon effect. A common strategy is to grow the tree until each node contains a small number of instances then use pruning to remove nodes that do not provide additional information. Pruning should reduce the size of a learning tree without reducing predictive accuracy as measured by a cross-validation set. There are many techniques for tree pruning that differ in the measurement that is used to optimize performance

REP (Reduced-Error Pruning) Tree refers to a tree created using the method is left to pruning some of the training set. When pruning a node is removed, subtree of the node under the node itself is leaf nodes. By using the validation set, it can be expected the effect of removing the added node by chance through the training set. REPTree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. It only sorts values for numeric attributes once. Missing values are dealt with using C4.5's method of using fractional instances.

3. SIMULATION

WEKA (The Waikato Environment for Knowledge Analysis) is a tool for data analysis and includes implementations of data pre-processing, classification, regression, clustering, association rules, and visualization by different algorithms. Implemented methods include instance-based learning algorithms, statistical learning like Bayes methods and tree-like algorithms like ID3 and J4.8 (slightly modified C4.5). Combinations of classifiers, e.g. bagging and boosting schemes, there are over sixty methods available in WEKA. WEKA is used as a tool for the experiment with the data iris.arff. It contains four properties and their values which are configuration of the data is sepallength, sepalwidth, petallength, petalwidth. The experiment described above was used for J48 and REPTree and the data analysis showed a 10-fold value of the Cross-Validation. k-fold Cross Validation is a way to ensure that there is no unique set of one share, compared with 'k'. With applying k-fold it is generated each algorithm of J48 and REPTree. The MAE (Mean Absolute Error) of J48 is 0.035, the classification accuracy of about 96% and MAE of REPTree is 0.0563, the classification accuracy of about 94% was confirmed by the experimental results shown in Table 3. In addition to analyzing the Kappa statistic metrics of recall (reproducibility) that indicates whether the experiment again similarly reproduced in if you run the same experiment at different times of the J48 0.94 average Kappa, Kappa is an average difference of 0.03 to about 0.91 of REPTree the show. Finally, means can check the RMSE (Root Mean Squared Error) is similar to the standard deviation of the statistically applying the J48, the average RMSE 0.1586, for REPTree showed the average RMSE 0.1936.

4. CONCLUSION

Data mining is widely used to obtain the useful information from each sector in company's customer management, the bank's personal and business credit score calculation, risk management, health care for the treatment of patients in clinical trials and DNA sequencing analysis in biotechnology. Also it is used in decision-making using variety of techniques for the diagnosis of patient's disease. In this paper, it is to determine the performance of J48 and REPTree algorithm for the same data in the experiment. Using the data to identify the varieties of irises as a percentage of the petal and a percentage of the sepal, a classification result of the iris according to the result of the petals was found the performance of the two algorithms. As a result, J48 classification success rate of approximately 96%, REPTree classification success rate of approximately 94%, J48 showed that the success rate of about 2% is excellent. As well as, Recall ratio in finely J48 was confirmed that the excellent performance. This allows better performance using the J48 algorithm for a more accurate prediction of the final target of the new data in the target and ML algorithms of this paper confirmed the results. Based on this, it is difficult to visually detailed classification and distinction. It is expected to be relatively difficult precise separation of many other breeds.

REFERENCES

[1] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[2] Baldi, Pierre, and Søren Brunak. Bioinformatics: the machine learning approach. MIT press, 2001.

[3] Kononenko, Igor. "Machine learning for medical diagnosis: history, state of the art and perspective." Artificial Intelligence in medicine 23.1 (2001): 89-109.

[4] Zhou, Zhi-Hua, and Min-Ling Zhang. "Solving multi-instance problems with classifier ensemble based on constructive clustering." Knowledge and Information Systems 11.2 (2007): 155-170.

[5] Bellazzi, Riccardo, and Blaz Zupan. "Predictive data mining in clinical medicine: current issues and guidelines." International journal of medical informatics 77.2 (2008): 81-97.

[6] Zhu, Xiaojin. "Semi-supervised learning literature survey." (2005).

[7] Cho, Sung-Bae, and Hong-Hee Won. "Machine learning in DNA microarray analysis for cancer classification." Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19. Australian Computer Society, Inc., 2003.

[8] Yongheng Zhao and Yanxia Zhang, Comparison of decision tree methods for finding active objects, Advances of Space Research, 2007

[9] D. L. Gupta, A. K. Malviya, Satyendra Singh Performance Analysis of Classification Tree Learning Algorithms, International Journal of Computer Applications (0975 – 8887) Volume 55– No.6, October 2012