APPLICATION WRAPPER-BASED FEATURE SELECTION ON C4.5 DECISION TREE CLASSIFIER

Jasmina Đ. Novakovic[1], Alempije Veljovic[2], Sinisa S. Ilic[3], Vladimir Veljovic[2]

[1] Belgrade business school, Higher education institution for applied science, Belgrade, Serbia
[2] Faculty of technical science Cacak, University of Kragujevac, Cacak, Serbia
[3] Faculty of technical science in Kosovska Mitrovica, University of Pristina, Kosovska Mitrovica, Serbia

Abstract

This paper presents the performance of C4.5 decision tree algorithm with wrapper-based feature selection. C4.5 decision tree has inherited ability to focus on relevant features and ignore irrelevant ones, but such method may also benefit from independent feature selection. Eighteen data sets were used for tests to compare results of classification accuracy with C4.5 decision tree algorithm. We proved that wrapper-based feature selection applied on C4.5 decision tree classifier effectively contributes to the detection and elimination of irrelevant, redundant data and noise in the data. In our experiments, wrapper-based feature selection improves classification accuracy of C4.5 decision tree algorithm.

Keywords: classification accuracy, C4.5 decision tree algorithm, J48, wrapper-based feature selection.

INTRODUCTION

Feature selection is a fundamental problem in many different areas. Feature selection reduces the dimensionality of feature space, removes redundant, irrelevant, or noisy data. Feature selection brings the immediate effects for application: speeding up a data mining algorithm, improving the data quality and the performance of data mining, and increasing the comprehensibility of the mining results.

All features may be important for some problems, but for some target concept, only a small subset of features is usually relevant. Finding the best feature subset is usually intractable [1] and many problem related to feature selection have been shown to be NP-hard [2].

Since 1970's feature selection has been a fertile field of research and development in statistical pattern recognition, machine learning and data mining [3]-[6].

Algorithms for feature selection may be divided into filters, wrappers and embedded approaches.

Some classification algorithms have inherited ability to focus on relevant features and ignore irrelevant ones. Decision trees are primary example of a class of such algorithms [7], [8], but also multi-layer perceptron (MLP) neural networks with strong regularization of the input layer may exclude the irrelevant features in an automatic way. But, some of these methods may also benefit from independent feature selection.

Section 2 presents in general the C4.5 decision tree algorithm. Section 3 describes the experiments and results. Section 4 concludes and gives future investigations.

2. C4.5 DECISION TREE

In machine learning different methods exist to build decision trees, but all of them summarize given training data in a tree structure, with each branch representing an association between feature values and a class label. Among these the C4.5 decision tree is one of the most famous and representative [9].

The C4.5 decision tree works by recursively partitioning the training data set according to tests on the potential of feature values in separating the classes. The decision tree is learned from a set of training examples

through an iterative process, of choosing a feature and splitting the given example set according to the values of that feature.

For this algorithm, the most important question is which of the features is the most influential in determining the classification and hence should be chosen first. Entropy measures or equivalently, information gains are used to select the most influential, which is intuitively deemed to be the feature of the lowest entropy (or of the highest information gain).

The C4.5 decision tree works by: a) computing the entropy measure for each feature, b) partitioning the set of examples according to the possible values of the feature that has the lowest entropy, and c) for each are used to estimate probabilities, in a way exactly the same as with the Naive Bayes approach. Although feature tests are chosen one at a time in a greedy manner, they are dependent on results of previous tests.

3. EXPERIMENTS RESEARCH AND RESULTS

Data sets taken from University of California, Irvine (UCI) repository of machine learning databases [10] were used for tests to compare results of classification accuracy with the J48, an open source Java implementation of the C4.5 decision tree algorithm.

We used these data sets: breast cancer (*bc*), credit approval (*ca*), Statlog german credit data (*cg*), cardiography (*ct*), hepatitis (*he*), liver (*li*), lung cancer (*lc*), mammographic mass (*ma*), monk problems (monk1 (*m1*), monk2 (*m2*), monk3 (*m3*)), mushrooms (*mu*), Parkinson (*pa*), Pima Indians diabetes (*pi*), image segmentation (*se*), soybean (*so*), Statlog heart (*sh*) and congressional voting records (*vo*).

Breast cancer (bc): The task of the data set is to predict whether or not there is recurrence of breast cancer. This data set includes 201 instances of one class (no recurrence of breast cancer), and 85 instances of another class (a recurrence of breast cancer).

Credit approval (ca): This file concerns credit card applications. This dataset is interesting because there is a good mix of attributes - continuous, nominal with small numbers of values, and nominal with larger numbers of values.

Statlog german credit data (*cg*): This dataset classifies people described by a set of attributes as good or bad credit risks. Comes in two formats (one all numeric).

Cardiography (ct): Data set consists of attribute measurement of fetal heart rate and uterine contractions attributes on ultrasound that are classified doctors [11]. This data set contains 2126 instances and 23 attributes.

Hepatitis (he): The main aim of data set is to predict whether hepatitis patients will die or not. There are two classes in data set: live (123 instances) and die (32 instances).

Liver (li): The first five variables in data set are all blood tests, which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each row in data set constitutes the record of a single male individual.

Lung cancer (lc): A set of data for the cancer of the lung contains data describing the three kinds of pathological forms of lung cancer. There are 32 instances and 56 attributes.

Mammographic mass (ma): The task is to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS features and the patient's age [12].

Monk problems (monk1 (*m1*), monk2 (*m2*), monk3 (*m3*)): There are three Monk's problems. The domains for all Monk's problems are the same. One of the Monk's problems has noise added.

Mushrooms (mu): This data set includes descriptions of mushrooms in terms of physical characteristics. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one.

Parkinson (pa): This data set consists of a range of biomedical voice measurements in 31 persons, 23 of them suffering from Parkinson's disease [13]. The main goal of this data set is to separate healthy people from those people who are suffering from Parkinson's.

Pima Indians diabetes (pi): In this data set the diagnostic is whether the patient shows

signs of diabetes according to World Health Organization criteria.

Image segmentation (*se*): The instances were drawn randomly from a database of 7 outdoor images. The images were hand segmented to create a classification for every pixel.

Soybean (so): This data set is Michalski's famous soybean disease database. There are 19 classes and 35 categorical attributes, some nominal and some ordered.

Statlog Heart (sh): The task is to predict absence or presence of heart disease. This data set contains 13 features (which have been extracted from a larger set of 74).

Congressional voting records (vo): This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA.

These data sets are data sets with a very large number of attributes, as well as those sets that have a small number of attributes which is good from the standpoint of research. Also, there are data sets that contain only categorical or numerical attributes, as well as data sets which contain both categorical and numerical attributes. Only two data sets have a greater number of classes, *se* with 7 classes, and *so* with 19 classes. The reason for this is the fact that in most problems of classification existing instances sort in two, possibly three classes, and rarely in a larger number of classes.

In this experiment WEKA (Waikato Environment for Knowledge Analysis) tools for data preparation and research are used. We used the classification accuracy as a measure of the quality of the model. We used cross-validation, because the procedure gives stable quality evaluation. The advantage of this method is that each of the *n* steps of cross validation using a large amount of data in their training and all available instances at one time was used to test. We take the value of *n* is 10.

Wrapper-based feature selection was used to reduce the dimensionality of data. The level of significance was set to a value of 0.05 in Paired *t*-test. Paired *t*-test was used to determine whether the value obtained by different methods differs significantly.

***TABLE I**. NUMBER OF ATTRIBUTES SELECTED BY THE WRAPPER-BASED FEATURE SELECTION*

Data set	Original	Wrapper
bc	9	3
ca	15	8
cg	20	9
ct	23	7
he	19	2
li	6	6
lc	56	2
ma	5	4
m1	6	3
m2	6	6
m3	6	3
mu	22	5
pa	23	5
pi	8	4
se	19	9
so	35	14
sh	13	3
vo	16	5

We compare two or more algorithms in our experimental research, and give a table of classification accuracy. The second algorithm is an algorithm in which was performed pre-selection attributes, and the first algorithm is a standard algorithm without pre-selection of attributes. In the table of classification mark "+" accuracy indicates a significantly higher value for classification accuracy, while "-" indicates a significantly lower value for classification accuracy.

The optimal number of attributes for the purposes of classification with wrapper-based feature selection is shown in Table I. Table I illustration the original size of the set and number of attributes selected by the wrapper-base feature selection.

Using wrapper-based feature selection, 14 data sets, from 18 observed, reduce the number of attributes exactly half or more than half compare with the original data set. The data set *lc*, with 56 attributes, has the greatest benefit of feature selection.

The accuracy of the classification algorithm using J48 with and without wrapper-based feature selection is shown in Table II and Figure 1.

TABLE II. J48 ALGORITHM AND CLASSIFICATION ACCURACY

Data set	J48	J48 reduced
bc	74.28	72.95
ca	85.57	84.43
cg	71.25	71.72
ct	98.57	98.88
he	79.22	81.90
li	65.84	66.36
lc	79.25	78.83
ma	82.19	82.47
m1	97.80	100.00
m2	63.48	65.72
m3	98.92	98.92
mu	100.00	100.00
pa	84.74	86.24
pi	74.49	73.44
se	96.79	96.73
so	91.78	91.74
sh	78.15	81.74
vo	96.57	95.24 -

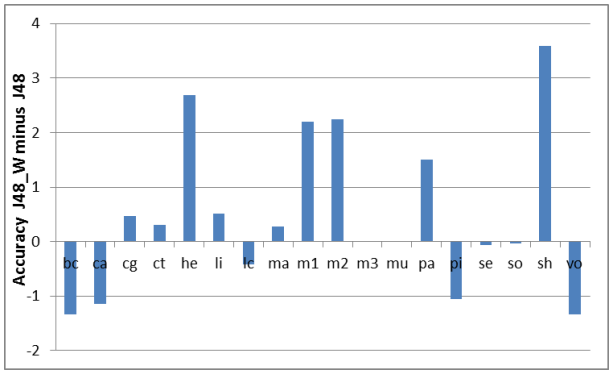

Fig. 1. *Classification accuracy*

We have significantly worse results of classification accuracy with wrapper-based feature selection (data set *vo*) in only one data set. We can conclude that the wrapper-based feature selection in most cases led to the same or better results in the observed data sets (in eleven data sets we have same or better results) with J48 classifier.

4. CONCLUSION

In experimental research we show it is possible to improve the classification accuracy of J48 algorithm, using the wrapper-based feature selection for reducing the dimensionality of the data. We implemented and empirically tested wrapper-based feature selection with J48 algorithm.

The wrapper-based feature selection effectively contributes to the detection and elimination irrelevant and redundant data, and also noise in the data. The wrapper-based feature selection selects relevant attributes and contributes to the greater classification accuracy in most data sets.

In further research we try to apply other techniques to solve the problem of dimensionality reduction of data and analyze and compare the effects of their implementation on J48 algorithm.

Acknowledgment

The authors are grateful for the support of the Ministry of Education, Science and Technological Development of Republic of Serbia - Projects TR 34009 and TR1653014.

REFERENCE

[1] R. Kohavi, G.H. John, "Wrappers for feature subset selection", *Artificial Intelligence*, vol. 97, 1997, 273-324.

[2] A.L. Blum, R.L. Rivest, "Training a 3-node neural networks is NP-complete", *Neural Networks*, 5:117–127, 1992.

[3] J. G. Dy, C. E. Brodley, "Feature subset selection and order identification for unsupervised learning", *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, 247–254.

[4] Y. Kim, W. Street, F. Menczer, "Feature selection for unsupervised learning via evolutionary search", *Proceedings of the Sixth ACM SIGKDD International Conference on knowledge Discovery and Data Mining*, 2000, 365–369.

[5] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection", *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[6] P. Mitra, C. A. Murthy, S. K. Pal, "Unsupervised feature selection using feature similarity", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):301–312, 2002.

[7] L. Breiman, J.H. Friedman, R.H. Olshen, C.J. Stone, Classification and regression trees, Wadsworth and Brooks, Monterey, CA, 1984.

[8] J.R. Quinlan, C4.5: Programs for machine learning, San Mateo, Morgan Kaufman, 1993.

[9] C. J. Merz, P. M. Murphy, UCI Repository of machine learning databases,

http://www.ics.uci.edu/˜mlearn/MLRepository.html, 1998.
[10] A. Frank, A. Asuncion, UCI Machine learning repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science, 2010.
[11] D. Ayres de Campos, et al., "SisPorto 2.0 A Program for automated analysis of cardiotocograms", *J Matern Fetal Med* 5:311-318, 2000.
[12] M.A. Little, P.E. McSharry, S.J. Roberts, D.A.E. Costello, I.M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection", *BioMedical Engineering OnLine*, 6:23, 2007.
[13] J. Đ. Novakovic, "Support Vector Machine as Feature Selection Method in Classifier Ensembles", *I.J. Modern Education and Computer Science*, 4, 1-8, 2014.