Using sensitivity as a method for ranking the test cases classified by binary decision trees

USING SENSITIVITY AS A METHOD FOR RANKING THE TEST CASES CLASSIFIED BY BINARY DECISION TREES

Sabrina Noblesse, Koen Vanhoof

Abstract: Usually, data mining projects that are based on decision trees for classifying test cases will use the probabilities provided by these decision trees for ranking classified test cases. We have a need for a better method for ranking test cases that have already been classified by a binary decision tree because these probabilities are not always accurate and reliable enough. A reason for this is that the probability estimates computed by existing decision tree algorithms are always the same for all the different cases in a particular leaf of the decision tree. This is only one reason why the probability estimates given by decision tree algorithms can not be used as an accurate means of deciding if a test case has been correctly classified. Isabelle Alvarez has proposed a new method that could be used to rank the test cases that were classified by a binary decision tree [Alvarez, 2004]. In this paper we will give the results of a comparison of different ranking methods that are based on the probability estimate, the sensitivity of a particular case or both.

ACM Classification Keywords: I.2.6 Learning – induction, concept learning; I.5.2 Classifier design and evaluation

1. Introduction

Decision trees do not only classify a particular case, they will also give the probability or the estimate of the probability that this case eventually belongs to the predicted class. For computing the probability estimate of a case, decision tree algorithms use the raw training frequency of the leaf where the particular case belongs to. The raw training frequency, P, puts the number of positive training cases, k, in a leaf to the total number of training cases, n, in that leaf [Zadrozny and Elkan, 2001]. If the class of the leaf of a particular training case is the same as the real class of that case, it is called a positive training case. The formula, $P = k / n$, can be used to compute the raw training frequency. This way of computing probability estimates and the fact that decision tree algorithms want to maximize classification accuracy and minimize the size of the decision tree [Provost and Domingos, 2000] cause some problems concerning the use of probability estimates for ranking the test cases classified by decision trees. These problems are listed below.

- Probability estimates can have extreme values: decision tree algorithms try to make the leaves of the decision tree as homogeneous as possible. By doing so the observed frequencies of positive training cases will shift automatically to 0 and 1 [Zadrozny and Elkan, 2001; Provost and Domingos, 2000].
- Probability estimates are not statistically reliable: this applies especially when the number of training cases associated with a leaf is small [Zadrozny and Elkan, 2001]. Suppose our decision tree has a leaf that consists of cases that all belong to the same class. It's difficult to accept that other cases also belong to the same class if they will comply with the constraints imposed by the tests of the decision tree that are needed to come to this particular leaf especially when only a small number of training cases belongs to this leaf [Provost and Domingos, 2000].
- One probability estimate per leaf: a decision tree algorithm assigns the same probability estimate to every case that belongs to the same leaf. We know that every leaf of a decision tree corresponds with a certain decision space. Thus, regardless of the fact that the different cases of a leaf are situated on different places in this space, a decision tree algorithm will appoint to all of them the same probability [Margineantu and Dietterich, 2001].

A possible solution for these problems is the topic of this paper. If the attribute values of a dataset are numeric and a binary decision tree is used to classify the test cases, one can use a sensitivity algorithm to compute the sensitivity or distance between a case and the corresponding decision surface [Alvarez, 2004]. Eventually the computed distance or sensitivity then can be used to rank the test cases.

This paper will compare the results of different ranking methods that are based on either the probability estimate, the sensitivity of a particular case or both. This way it will be possible to find out if the sensitivity algorithm can be used for ranking the test cases classified by a binary decision tree.

2. Experimental Design

This experiment was conducted with the dataset of the data mining competition that took place in the run up to the 29th Annual Conference of the German Classification Society (GFKL 2005): From Data Analysis to Knowledge Engineering. This dataset was provided by the sponsor Deutsche Sparkassen und Giroverband (DSGV) and the goal of the competition was to predict a liquidity crisis based on a subset of 26 variables describing attributes of companies. When the companies were facing a liquidity crisis they should be classified as belonging to class 1 and if they don't have a liquidity crisis they belong to class 0. Only 11% of the data set has a liquidity crisis. Because the values of these attributes were all numeric, this dataset could be used to test the sensitivity algorithm of Isabelle Alvarez and compare its results with the results of other ranking methods. In order to compare the different ranking methods we used all attributes to select the 2000 companies out of 10000 companies from the test dataset that were most likely to be correctly classified according to the used ranking method. The number of correctly and incorrectly classified cases of the test dataset will then be compared for the different ranking methods so we can draw some conclusions about the use of the sensitivity algorithm.

We used the Weka j48 algorithm [Witten and Frank, 2000] to grow a binary decision tree on the training dataset. From the resulting model the different decision surfaces could be derived. A decision surface can be seen as the boundary between regions with different class labels. Every leaf within the tree has its own decision surface determined by the tests a case has to comply with before belonging to a particular leaf. Because the emphasis of this study lays on the examination of the sensitivity algorithm as a ranking method and not on the development of a state of the art decision model/decision tree, we used the default values for the different thresholds and parameters that are necessary for developing a binary decision tree. Once Weka had developed the binary decision tree it could be applied to the test dataset and after predicting the class value for each case of the test dataset, we were able to apply the earlier mentioned sensitivity algorithm. Two kinds of standardization methods were applied on the attributes, the standard and the minmax method. Both methods are defined with information that could be easily inferred from the dataset itself.

$$y_i^{MinMax} = (x_i - Min_i) / (Max_i - Min_i)$$

$$\text{or} \quad y_i^{Standard} = (x_i - E_i) / S_i$$

with min, max, E,S respectively the minimum, the maximum, the estimated mean value and standard deviation of the corresponding attribute values. The sensitivity algorithm projects a given case onto the decision surface of every leaf that has a different class value than the class value predicted for this case. The algorithm then computes and ranks the distance between the case itself and its different projections. We assume that when the distance between a case and the decision surface of a leaf becomes smaller, the risk of a misclassified instance becomes more realistic. The final step of this experiment is to rank the classified instances by different ranking methods based on the probability estimates and/or sensitivity and compare the results.

3 Methods for Ranking Test Cases Classified by Binary Decision Trees and Experimental Results

In this section, we present the methods we used for ranking the test cases that were classified by a binary decision tree. More specifically we will look at some ranking methods and give in section 4 their impact on selection and ranking. The impact of selection is done by analyzing a) the number of correctly classified cases, b) the number of selected minority cases and c) the correct classification of minority cases. The impact on the ranking is done by analyzing the Roc curves. These ranking methods are based on the probability estimates, the sensitivity of a case or a combination of both. For comparing the different ranking methods we will look at the 2000 best cases. The ranking methods are based on:

- Transformed probability estimate
- Sensitivity
- Piecewise correction function

3.1 Transformed Confidence or Probability Estimates

Within the scope of the data mining competition it was important to correctly classify as much cases of class 1 as possible. These are the cases that either belong to the True Positive quadrant or to the False Negative quadrant. We assume that the cases that belong to the False Negative quadrant are the ones that were classified as class 0 but have a very low probability estimate. To be able to select also these cases we can try to convert the probability estimates of the cases that belong to class 0 to the probability that they would belong to class 1. This can be done with the following formula.

$$P^*0 = 1 - P0$$

With P0 the probability that the case belongs to class 0 and P*0 the transformed probability estimate. We can use this formula for transforming the probability estimates for cases belonging to class 0 because the classified cases can only belong to class 0 or class 1. This is why we can assume that if the case doesn't belong to class 0, it can only belong to class 1. The probability estimates of the cases belonging to class 1 will not be transformed.

3.2 Sensitivity

The sensitivity of a case from the dataset corresponds with the smallest distance between the case and the corresponding decision surface with different class value. This surface is created by the decision tree. We assume that how smaller the distance, how larger the probability of an incorrect prediction or a wrong classification. For computing the sensitivity we have used the decision surfaces from the decision tree that was developed for the data mining competition. The distance between the classified cases and the corresponding decision surfaces was then computed. Like already stated we are using two different standardization methods for calculating the sensitivity, the Standard and the Minmax method. The Minmax method makes the attribute values laying between 0 and 1 and the standard method standardizes the different values of the attributes.

3.3 Piecewise Correction Function

In the following, we will describe a function that can be used for "correcting" the probability estimates made by a binary decision tree. Because the resulting ranking method has to take into account both the probability estimate and the sensitivity of the concerning/particular case, both factors need to be present in the equation of the correction function. After interpreting the definition of sensitivity and the initial goal of the sensitivity algorithm of Isabelle Alvarez [Alvarez, 2004] we believe that when the sensitivity or the distance from a case to the decision surface is small, the probability of an incorrect prediction or classification will be greater. When the distance is larger there would be, according to us, less chance on an incorrect prediction or classification by the developed decision tree. We assumed that there could be a different correction necessary for a sensitivity value that is below

a certain threshold than for a probability estimate whose value is above a specific threshold. Keeping all this in mind we have chosen for a piecewise correction function.

The modified probability estimate will be presented by the symbol *W** and the initial probability estimate by the symbol *W*. We use the symbols *s* and s_0 to indicate the sensitivity and the threshold of the sensitivity. The symbols c_1 and c_2 denote the amount by which the original probability estimate will be altered. The following piecewise function could be used for correcting the original probability estimates.

$$W^* = W + c_1 \text{ if } s > s_0$$
$$W^* = W - c_2 \text{ if } s \leq s_0$$

Note that the modified estimate can be greater then one or less than zero and cannot be considered any more as a probability estimate. After the selection of a threshold s_0 we will examine for this threshold what the effects of different corrections of the probability estimates will be on the results of the ranking. The extent by which the original probability estimates will be corrected varies between 0 and 50 percent. The choice for a particular value for the parameters c_1, c_2 and s_0 will depend on the effect of the ranking of the training data set. A linear search has been applied to maximize the effects on the training data.

The Minmax metric and the Standard metric were treated separately when developing a correction function for the probability estimates given by the binary decision tree because they have for each case a different sensitivity and both metrics also have different maximum sensitivity values. The sensitivity that was computed with the Standard method lies in the interval [0 ; 1,375939965248] and in the interval [0 ; 0,0235140007] for the Minmax method.

4 Selection and Ranking Results

The evaluation of the results from the different ranking methods will be done by comparing the 2000 most likely cases of every applied ranking method with the solution dataset sent to us by the organizers of the data mining competition.

4.1 Selection

The individual selection results of each ranking method will be extracted from the corresponding confusion matrix. The results will be put in a tabular overview as can be seen in Table 1. A confusion matrix shows the number of cases that are correctly classified as class 0 and class 1 and the cases that are incorrectly classified as class 0 and class 1. Because we have only two possible classes in our experiment, the confusion matrix can be depicted as a 2x2-matrix. The classifications that end up in the True Positive and the True Negative quadrant are correctly made classifications, these are cases that were classified as class 0 and class 1 and also really belong to this classes. A case that lies in the False Positive quadrant is a case that is a class 0 case in reality but was classified as class 1. If a case was classified as class 0 but actually belongs to class 1, it will be placed in the False Negative quadrant.

Our quality criteria are

a) TP +TN : the total accuracy of classified cases,
b) TP +FN : the number of selected minority (positive) cases,
c) TP/ (FP+TP) : the accuracy of minority (positive) case classification.

The first criterion is important when both class values are of equal importance. So the bank can handle companies with our without a liquidity crisis in an appropriate way. This is called the accuracy criterion. The second criterion is important when the bank wants to reach all companies with a liquidity crisis (positive case) and it does not matter when companies without a liquidity crisis are also included. This is called the market share criterion. The third criterion is important when the bank wants to reach the companies with a liquidity crisis (positive case) but without reaching companies without a liquidity crisis. This is called the profit share criterion. Next table gives the results.

Table 1: Results of the selection of 2000 test cases using the different ranking methods

Ranking method	Criteria		
	Total Accuracy	Number of minority cases	Accuracy Minority
Transformed probability estimates	70.55	753	0.62
Sensitivity			
Standard metric	89.65	495	0.70
Minmax metric	86.50	501	0.68
Correction function			
Standard metric	79.3	753	0.63
Minmax metric	71.3	753	0.63

From this table we learn that indeed using sensitivity increases the accuracy of the classifier and therefore selects less minority cases (= most difficult to predict). When the goal of the company is to maximize profit or accuracy : the ranking method based on sensitivity is the best method. When the goal of the company is market share : the ranking method based on transformed probability is the best method. Applying a correction function delivers a compromise solution.

4.2 Ranking

To be able to compare the ranking performance of different ranking methods, a single number measure which reflects the ranking performance of the ranking method is needed. The area under the ROC curve (AUC) appears to be one of the best ways [Bradley A.P.,1997], [Ling, C. X., Huang, J. and Zhang, H., 2003]. This ROC space is a coordinate system where the rate of true positives is plotted on the Y-axis and the rate of false positives is plotted on the X-axis for every proportion of the ranked dataset. The true positive rate is defined as the fraction of positive cases classified correctly relative to the total number of positive examples. The false positive rate is defined as the fraction of negative cases classified erroneously relative to the number of all negative examples. From a visual perspective, one point in the ROC curve is better than another if it is located more to the north-west (TP is higher, FP is lower or both) on the ROC graph. Statistical analysis was applied to calculate upper and lower bounds (98% confidence level. The results are given in next table.

Table 2: Results of the Roc analysis using the different ranking methods

Ranking method	Criteria		
	Area under curve	Lower bound	Upper bound
Transformed probability estimates	0.731	0.709	0.754
Sensitivity			
Standard metric	0.887	0.865	0.908
Minmax metric	0.808	0.781	0.834
Correction function			
Standard metric	0.841	0.822	0.861
Minmax metric	0.747	0.725	0.769

These results confirm previous findings. With sensitivity we can better predict the minority cases and as a consequence in the ROC graph these cases are ranked first and a higher “Area under curve” will be obtained as can be seen by comparing the left and right part of figure 1.

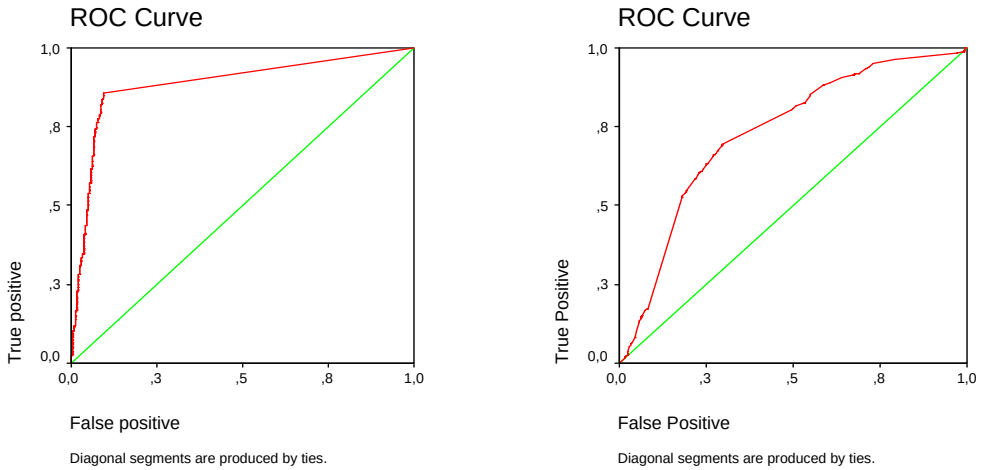

Figure 1: Roc Curve for standard metric (left) and for transformed probability (right)

It is clear that for the sensitivity ranking methods the difference is quite large and statistically significant. The confidence intervals do not overlap with the confidence intervals of transformed probability.

This figure clearly show that the use of probability estimates for ranking the test cases classified by decision trees can be improved by using sensitivity.

5 Conclusion

Based on the comparison of the ranking methods in the previous section, it has become clear that business goal (accuracy, profit or market share) and the way we rank the classified cases will determine the final results. The ranking methods based on sensitivity give the highest total accuracies, the best profit and give the highest Area under curve figures. If we use a ranking method that is based on the sensitivity then the used standardization method will also be of importance for the results. In our experiment the results are optimal when it is based on computations with the Standard metric. The mentioned problems concerning the use of probability estimates for ranking the test cases classified by decision trees are confirmed. When the task is selection of minority cases the position of the cut-off point (2000 out of 10 000 in our case) is also of crucial importance. It seems that decision trees perform well when the cut-off point is far away from the number of minority cases (1113).

Finally, we want to mention that the results of this experiment are valid for the data set under study and can not yet be generalized because further research will be necessary to decide if using the sensitivity of the cases in a dataset is a good basis for the correction of probability estimates of binary decision trees or for the improvement of ranking the results of binary decision trees.

Bibliography

[Alvarez, 2004] I.Alvarez. Sensitivity Analysis of the Result in Binary Decision Trees. In: Proceedings of the 15th European Conference on Machine Learning, Vol 3201, pp 51-62, Springer Verslag.

[Bradley A.P.,1997] The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognition 1997; 30 (7): 1145-1159.

[Ling, C. X., Huang, J. and Zhang, H., 2003] AUC: a statistically consistent and more discriminating measure than accuracy. In Proceedings of 18th International Conference on Artificial Intelligence (IJCAI-2003), 2003. p. 329–341.

[Margineantu and Dietterich, 2001] D.D.Margineantu and T.G. Diettterich. Improved Class Probability Estimates from Decision Tree Models. Departement of Computer Science, Oregon State University.

[Provost and Domingos, 2000] F. Provost and P. Domingos. Well-Trained PETs: Improving Probability Estimation Trees. Information Systems Department, Stern School of Business, New York University.

[Witten and Frank, 2000] I. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation. Morgan Kaufmann.

[Zadrozny and Elkan, 2001] B. Zadrozny en C. Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. Department of computer science and engineering, University of California.

Authors' Information

Prof. Dr. Koen Vanhoof – Universiteit Hasselt, Campus Diepenbeek, Departement BEDR/VEBI; Agoralaan gebouw D, B3590 Diepenbeek, Belgium; e-mail: koen.vanhoof@uhasselt.be

Sabrina Noblesse – Universiteit Hasselt, Campus Diepenbeek, Departement BEDR/VEBI; Agoralaan gebouw D, B3590 Diepenbeek, Belgium

THE STAPLE COMMODITIES OF THE KNOWLEDGE MARKET

Krassimir Markov, Krassimira Ivanova, Ilia Mitov

Abstract: In this paper, the "Information Market" is introduced as a payable information exchange and based on it information interaction. In addition, special kind of Information Markets - the Knowledge Markets are outlined. The main focus of the paper is concentrated on the investigation of the staple commodities of the knowledge markets. They are introduced as kind of information objects, called "knowledge information objects". The main theirs distinctive characteristic is that they contain information models, which concern sets of information models and interconnections between them.

Keywords: Information Market, Knowledge Market, Knowledge Information Objects, General Information Theory

ACM Classification Keywords: K.4 Computers and Society – K.4.0 General; K.4.4 Electronic Commerce

> "The speaker doesn't deliver his thought to the listener, but his sounds and performances provoke the thought of the listener. Between them performs a process like lighting the candle, where the flame of the first candle is not transmitted to another flame, but only cause it."
>
> *Pencho Slaveikov, Bulgarian poet,*
> *the beginning of the XX-th century*

Introduction

The main characteristic of the Information Markets is payable information exchange and based on it information interaction. Special kinds of Information Markets are the Knowledge Markets. The main goal of this paper is to continue the investigation of the Knowledge Markets started in [Ivanova et al, 2001], [Markov et al, 2002]. Now, our attention will be paid to the staple commodities of the Knowledge Markets. The usual talk is that at the Knowledge Market one can buy knowledge. But, from our point of view, this is not so correct.

The investigation presented in this paper is based on the *Theory of Information Interaction*, which is one of the main parts of the *General Information Theory* [Markov, 1984], [Markov, 1988], [Markov et al, 1993], [Markov et al, 2003].