Predicting Customer Behavior using Naive Bayes and Maximum Entropy – Winning the Data-Mining-Cup 2004 –

Arne Mauser, Ilja Bezrukov, Thomas Deselaers, Daniel Keysers
Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{mauser,bezrukov,deselaers,keysers}@i6.informatik.rwth-aachen.de

Abstract: In this work we describe combinations of classifiers using Naive Bayes, Maximum Entropy, Neural Networks and Logistic Regression for classification of customer records. Performance of these approaches is confirmed by the 1st, 3rd, and 5th rank in the Data-Mining-Cup 2004.

1 Introduction

With the increasing possibility of collecting data in business applications there is a rising demand to utilize the available information. For financial institutions it may be the detection of fraudulent transactions or prediction of a company's liquidity situation. In sales it might be the prediction of any aspect of customer behavior.

A major expense factor in mail-order business is the cost of returns. In 2004, prediction of returning behavior of customers of a major German mail-order company was the task of the Data-Mining-Cup competition. The Data-Mining-Cup is an annual student competition organized by the Chemnitz University of Technology and Prudsys AG.

In this paper, we are going to present the approaches, that obtained the 1st, 3rd, and 5th rank out of 97 participants. A Maximum Entropy combination of Naive Bayes classifiers won the Data-Mining-Cup in 2004. Combining Logistic Regression, Neural Networks, and Maximum Entropy obtained the 3rd rank.

2 Task description & approach

The data, provided by a German mail-order company was split into a test set and a training set. Each consisted of 20147 customer records with 65 attributes. The attributes contained information about the customer's ordering behavior (e.g. "value of ordered goods in period C"), as well as geographical and statistical information (e.g. "fraction of households with children in the customer's ZIP-Code area").

Customer records were separated into 3 different classes, each class representing a different returning behavior. Customers who returned less than 18% of ordered goods were classified as “low returners” (L), those customers returning more than 40% were classified as “high returners” (H), those in between were labelled as “indefinite” (I).

For the test data the class labels were unknown and had to be predicted. A cost matrix for weighting classification errors between classes was used within the evaluation. Correct predictions of high and low returners were assessed with +1 points, wrong predictions with -1 points. With indefinite returners, correct classifications were rewarded with +0.5 points while errors were neglected. Each classification was assessed according to this measure and the result was summed over all records resulting in a global score.

3 Data preprocessing

Real-world data usually suffers from deficiencies that make classification tasks difficult: missing values, outliers, and noisy distributions affect the performance of classification algorithms. Furthermore, for many classifiers it is necessary to adjust feature values to a common interval. Generalization and noise reduction through histograms may also help to improve classification.

In our approach, the majority of features were transformed into “equi-depth” histograms. Bin borders were adjusted so that each bin contained approximately the same number of elements. Each individual feature value was then replaced by the center of the bin it belonged to. Using 5-fold crossvalidation, we determined that a 10-bin histogram performed best on the training data.

For features that contained a percentage of returned goods, we additionally generated two binary features for zero and missing values. The remaining features, like the ZIP-Code or the title of the customer were replaced with binary features. The decisions on what transformation to apply to the individual features were taken manually, based on the feature description.

4 Data classification

In order to estimate the performance of classification algorithms on unseen data and to find the most suitable approach, we evaluated several methods on the training data using 5-fold crossvalidation.

The four classification algorithms we tested performed similarly well: Logistic regression [HK00], Neural Networks [Bis96], Naive Bayes [HK00] and Maximum Entropy [BPP96]. In the following paragraphs, we will give a brief description of the latter two methods.

Naive Bayes classifiers rely on the bayesian decision rule with the assumption that the probability of each attribute value is independent from the values of other attributes. This assumption is not valid in general, but provides the advantage of computational simplicity.

Table 1: Results of classification methods obtained using 5-fold crossvalidation on training data.

Method	Error rate	Score
Logistic regression	27.2	10539.0
Neural Network	27.3	10496.5
Maximum Entropy	27.0	10610.0
Naive Bayes	39.0	5774.0
Logistic Regression + Neural Network + Maximum Entropy	26.9	10664.0
Naive Bayes + Maximum Entropy	27.0	10613.0

Maximum Entropy classifiers model the marginal distributions of the training data while trying to be as general as possible. Among the set of possible models the one with the highest entropy is chosen in training.

Combining different classifiers often improves the results over using a single classifier as disadvantages of one method might be compensated by others.

We used two methods for classifier combination. In our first approach we combined separately trained classifiers using the sum rule. That is, we summed up the a posteriori probabilities of the classifiers and chose the class having the maximum sum. Using this method, we combined Logistic Regression, Neural Network and Maximum Entropy classifiers.

Our second approach to classifier combination used a Naive Bayes classifier. No assumption was made regarding the distribution of the values. Probabilities were estimated by relative frequencies. Unseen values in test were replaced by their nearest neighbors. Instead of taking the product of the models estimated for the individual features, we used a Maximum Entropy approach for weighting the individual distributions.

For convenience, we reduced the number of classes to two, neglecting the class of "indefinite" returners, as this class was strongly underrepresented. Sharing the probability mass of the neglected class equally between the remaining two other classes, we used the probabilities of the Naive Bayes distributions for class "L" as feature functions. For training the feature function weights we used the Generalized Iterative Scaling algorithm as described in [KON02].

5 Results

In Table 1, we summarize the results for the described classification methods using 5-fold crossvalidation on training data. For the competition, parameter estimation was performed using all available training data. Table 2 shows the final scores of the top 5 submissions on the unseen test data as determined by the competition organizers.

The combination of Naive Bayes and Maximum Entropy scored best on the test set, although it did not perform as well on the training set. This can be explained by a slight over-fitting to the training data of our first combination approach.

Table 2: Top 5 submissions in the competition.

	University	Methods used	Score
1.	RWTH-Aachen	Naive Bayes + Maximum Entropy	10511.0
2.	Warsaw University	Support Vector Machine	10491.0
3.	RWTH-Aachen	Log. Reg. + Neural N. + Max.Ent	10490.0
4.	Handelshochschule Leipzig	Unknown	10459.0
5.	RWTH-Aachen	Maximum Entropy	10455.0

6 Conclusion

Using histogramization as data preprocessing steps and various classifier combinations we obtained excellent results in the Data-Mining-Cup 2004.

It can be seen that our data transformation was a good choice for the given task, as different classifiers perform equally well on this data. A Maximum Entropy combination of a set of Naive Bayes classifiers performed best on the unseen test data. A combination of Logistic Regression, Neural Networks, and Maximum Entropy using the sum rule performed almost as well. This shows that, given a set of well transformed data, the selection of the individual classifier may be less important.

Furthermore we observe that it is essential to handle the issue of over-fitting with care.

References

[Bis96] C. M. Bishop. *Neural networks for pattern recognition*. Clarendon Press, 1996.

[BPP96] A. L. Berger, S. Della Pietra, and V. J. Della Pietra. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71, 1996.

[HK00] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 1st edition, 2000.

[KON02] D. Keysers, F. J. Och, and H. Ney. Maximum Entropy and Gaussian Models for Image Object Recognition. In *Pattern Recognition, 24th DAGM Symposium, Zürich, Switzerland*, volume LNCS 2449 of *Lecture Notes in Computer Science*, pages 498–506. Springer-Verlag, 2002.