Hollmén,J.,M.Skubacz,and M.Taniguchi (2000). Input dependent misclassification costs for cost-sensitive classifiers. In N.Ebecken and C.Brebbia (Eds.), DATA MINING II - Proceedings of the Second International Conference on Data Mining 2000, pp.495-503. Published by WIT Press, Ashurst Lodge, Southampton, UK. http://www.witpress.com

Input dependent misclassification costs for cost-sensitive classifiers

J. Hollmén[1], M. Skubacz[2] & M. Taniguchi[2]
[1] *Laboratory of Computer and Information Science, Helsinki University of Technology, Finland*
[2] *Information and Communications, Corporate Technology, Siemens AG, Germany*

Abstract

In data mining and in classification specifically, cost issues have been undervalued for a long time, although they are of crucial importance in real-world applications. Recently, however, cost issues have received growing attention, see for example [1,2,3]. Cost-sensitive classifiers are usually based on the assumption of constant misclassification costs between given classes, that is, the cost incurred when an object of class j is erroneously classified as belonging to class i. In many domains, the same type of error may have differing costs due to particular characteristics of objects to be classified. For example, loss caused by misclassifying credit card abuse as normal usage is dependent on the amount of uncollectible credit involved. In this paper, we extend the concept of misclassification costs to include the influence of the input data to be classified. Instead of a fixed misclassification cost matrix, we now have a misclassification cost matrix of functions, separately evaluated for each object to be classified. We formulate the conditional risk for this new approach and relate it to the fixed misclassification cost case. As an illustration, experiments in the telecommunications fraud domain are used, where the costs are naturally data-dependent due to the connection-based nature of telephone tariffs. Posterior probabilities from a hidden Markov model are used in classification, although the described cost model is applicable with other methods such as neural networks or probabilistic networks.

1 Introduction

Classification has important applications in data mining. Having modeled the characteristics of classes, we can use Bayes rule for making statements about class membership of data, on which decisions can be based on. Optimal decisions are dictated by decision goals, which are specific for a given problem. Automatic classification is important in practical applications, like in medical diagnosis and fraud detection. Such domains involve large amounts of data, which is cumbersome and expensive to analyze by human experts. Consequences of decisions taken by a classifier are another important aspect of such domains. In cancer screening, for instance, a wrong decision has far reaching implications. Considering a healthy patient classified as sick or a sick patient as healthy are very different in nature. Furthermore, an incorrectly classified sick patient suffers from an uncured disease that may lead to death, but a misclassified healthy patient is only subject to additional examination before the correct decision. In fraud detection, the timeliness of correct decisions has a direct connection to monetary loss. On the other hand, a fraud detection system is faced with the annoyance of unnecessarily interrogated customers.

According to [4], cost issues are discussed surprisingly little in the literature. Often, they are neglected entirely and if considered, the standard way of incorporating them in classification is to state a fixed misclassification cost matrix, which summarizes the cost of misclassifying an entity [5]. Such statements do not include specific characteristics of a given entity, but are based on the average costs for the entire population. Additionally, cost structure may be misused to tackle the problem of highly imbalanced class distributions. For this approach, fixed misclassification cost matrix is sufficient, since in cost-aware decisions it has the same effect as corrected prior probabilities. This is equivalent to replacing the common with the cheap and the rare with the expensive. Although this assertion often happens to be successful, the very problem of real costs is not addressed. In this paper, we introduce a cost model that incorporates the specific properties of objects to be classified. Instead of a fixed misclassification cost matrix, we now have a more general matrix of cost functions. These functions operate on the data to be classified and are re-calculated for each data point separately. Using telecommunication fraud as an example, we incorporate the real cost of telephone calls made in our decisions. In [2], use of a cost model in evaluation and fine-tuning of a fraud model is reported, where the cost reflects the summed connection time of mobile phone calls. Our work, in contrast, integrates a similar cost model in the classification phase itself. In section 2, we review the Bayes rule in classification and relate our new method (section 2.3) to the readily established methods of cost-ignorant (section 2.1) and cost-aware classification (section 2.2). In section 3, experiments in the telecommunications domain are presented. The paper proceeds with a discussion in section 4 and ends with a summary in section 5.

2 Decision Goals and Classifier Design

The Bayes rule forms the foundation of pattern recognition and embodies the definition of conditional probability. The main application of it is to invert the class conditional probabilities of data to data conditional probabilities of classes. In the following equation, we denote class i by ω_i and data by x.

$$P(\omega_i, x) = P(\omega_i|x)P(x) \; ; \; P(x, \omega_i) = P(x|\omega_i)P(\omega_i)$$

By symmetry, we equate the above equations and solve for $P(\omega_i|x)$ to get the Bayes rule

$$P(\omega_i|x) = \frac{P(\omega_i)P(x|\omega_i)}{P(x)}.$$

The term $P(\omega_i)$ is the prior class probability of the class i, or the general knowledge of the class prevalence. The term $P(x|\omega_i)$ expresses the likelihood of data x under the assumption of class i. The denominator serves merely as a normalization factor.

Many classifiers use the above mentioned probabilities as the basis for decision-making. Before put to use, however, a decision goal must be formulated. Various decision goals come from different setups of the problems and lead to the corresponding decision functions. For instance, the goal to minimize the probability of misclassification leads to classifying samples to the class with the largest posterior probability [5]. This is valid for problems, where classes have equal importance. When highly imbalanced class distributions are encountered and the rare class is the interesting one, more advanced solutions are called for.

2.1 Equal Misclassification Cost

Cost-neutral approach to classification assumes equal misclassification costs between classes. This decision goal is realized through the so-called maximum posterior classification rule, in which a sample is classified to the class with the largest posterior probability [5]. Imbalanced priors are often encountered in practical problems, like medical screening or fraud detection. Expressing negative log posterior for the class ω_2 as in the equation below, we can see the leverage effect of the ratio of class priors $\frac{P(\omega_1)}{P(\omega_2)}$ on the ratio of class likelihoods $\frac{P(x|\omega_1)}{P(x|\omega_2)}$

$$-\log P(\omega_2|x) = \log\left(1 + \frac{P(\omega_1)P(x|\omega_1)}{P(\omega_2)P(x|\omega_2)}\right).$$

2.2 Fixed Misclassification Cost

Neglecting cost issues is unacceptable in many domains. The standard approach to incorporating costs in decision-making is to define fixed and unequal misclassification costs between classes. Cost model takes the form of a cost matrix, where the

cost of classifying a sample from a true class j to class i corresponds to the matrix entry λ_{ij}. This matrix is usually expressed in terms of average misclassification costs for the problem. The diagonal elements are usually set to zero, meaning correct classification has no cost. We may define conditional risk for making a decision α_i as

$$R(\alpha_i|x) = \sum_{j=1}^{n} \lambda_{ij} P(\omega_j|x).$$

The equation states that the risk of choosing class i is defined by fixed misclassification costs and the uncertainty of our knowledge about the true class of x expressed by the posterior probabilities. The goal in cost-sensitive classification is to minimize the cost of misclassification, which can be realized by choosing the class with the minimum conditional risk. The decision function for the two-class case becomes then

$$R(\alpha_1|x) \underset{\alpha_2}{\overset{\alpha_1}{\lessgtr}} R(\alpha_2|x).$$

This notation means that action α_1 is chosen if the left-side of the inequality is smaller than the right-side of it. Action α_2 is chosen, if the opposite holds. Expressing the decision function as a posterior probability of the class ω_2 we get a linear function according to which decision threshold is selected.

$$[\lambda_{12} + \lambda_{21}]P(\omega_2|x) - \lambda_{21} \underset{\alpha_1}{\overset{\alpha_2}{\gtrless}} 0$$

2.3 Input-dependent Misclassification Cost

In practice, the same type of misclassification may have different costs depending on the object to be classified, contrary to the fixed misclassification cost approach, where costs remain constant regardless of the data to be classified. Costs are often naturally derived from the problem setting and may involve transaction costs or other factors involving direct monetary loss. Using this approach, the conditional risk for making a decision α_i is defined as

$$R(\alpha_i|x) = \sum_{j=1}^{n} \lambda_{ij}(x) P(\omega_j|x)$$

where the $\lambda_{ij}(x)$ is the misclassification cost function taking into account the properties of the data point x, for example the amount of credit used, and is recalculated for each case separately. Typically, these functions are not very complex, but are naturally defined by the real-life setting of costs.

The form of the decision function, which minimizes the conditional risk of our action remains the same as in the previous section, only the cost depends now on

the input data. The risk minimizing decision function is

$$[\lambda_{12}(x) + \lambda_{21}(x)]P(\omega_2|x) - \lambda_{21}(x) \underset{\alpha_1}{\overset{\alpha_2}{\gtrless}} 0.$$

This is similar to the decision function presented in the previous section, except that the factors, which were constant are now data-dependent. As a consequence, decision function becomes non-linear. In the next section on experiments, we will illustrate and discuss our new approach more closely and formulate the cost model for our fraud detection problem.

3 Experiments

In order to compare the effectiveness of previously described methods, we present empirical experiments in the telecommunications fraud domain. The summed length of calls of one day is used as an observed variable in a hidden Markov model (HMM) [6]. The experiments are performed on real data involving fraudulent behavior and on similar, simulated data obeying theoretical assumptions.

For assessment, we use a cost model that measures the gross profit of the operator. This model should reflect the practical costs involved in the operation of a real network. Therefore, cost of fraudulent activity is treated as lost revenue. Moreover, in the case of legitimate subscribers falsely classified as fraudsters we discount the future profit by a constant factor, e.g. 0.99. This accounts for annoyance on the customer's part due to an interrogation. Additionally, a fixed transaction cost is calculated for needless inquiries. This cost model is an extension of that presented in [2]. The parameters for the profit calculation are presented in the appendix.

To model the distributions of normal and fraudulent populations, we use a hidden Markov model [6]. In HMM, the observed variables are assumed to be conditionally dependent on a discrete hidden variable. The hidden variable undergoes a transition in time according to a linear transition matrix. In our models, we assume two hidden states $s_i, i = 1,2$ for modeling legitimate and fraudulent subscribers. The probability densities for the observed variables are $P(x|s_i) = \lambda_{s_i} e^{-\lambda_{s_i} x}$. This assumes the summed length of calls of the current day to be exponentially distributed. The transition matrix was set to reflect the real dynamics of fraud. More specifically, the expected time spent in the fraud state was set to be shorter than that spent in the normal state and the probability of entering the fraud state was set to be low. We furthermore extend the HMM to include the effect of imbalanced population priors by setting a prior for the hidden variable. Through inference, we calculate the time-varying posteriors for normal and fraudulent behavior, on which the detection is based. The parameters of our model are given in the appendix. Real data used in the experiments spanned a period of seven weeks. Activity of 304 fraudulent and 1988 legitimate users was recorded. For each subscriber, we

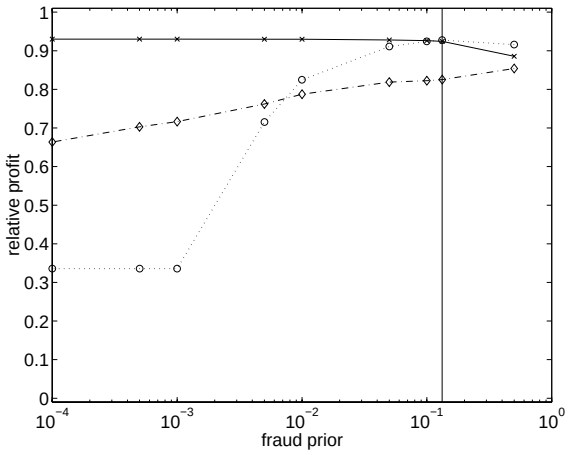

Figure 1: The results from the real data experiment.

calculated the total length of calls performed during one day and used it as an observed variable in our HMM. Since the priors are unknown in practical problems, we performed a series of experiments assuming different values for priors. For each setting of priors, we compared the profit of the operator using fraud detection with equal, fixed and input-dependent misclassification cost approaches based on the same class posteriors. Fixed misclassification cost for false negatives was calculated as the inverse of the assumed prior, the false positive cost was unity. For the input-dependent misclassification, the cost for a false negative is $\lambda_{12}(x) = ux$, where x is the summed length of calls of the current day and for the false positive $\lambda_{21}(x) = a + (1-k)ux$. The value of $k = 0.999$ corresponds to discounting the profit due to dissatisfaction of the customer and $a = 1$ is a transaction cost. The pricing of calls assumes unit price $u = 1$ for one unit of air time. The input dependent misclassification cost matrix Λ is

$$\Lambda = (\lambda_{ij}) = \begin{pmatrix} 0 & ux \\ a + (1-k)ux & 0 \end{pmatrix}.$$

The same parameter values were used in the evaluation phase. The same set of experiments with artificial data was repeated to gain additional control over the experimental setting. In this case, the same model from which the data was sampled was also used in the detection phase in order to rule out any inaccuracies in the model estimation. We sampled 100 fraudulent users and 10000 normal users from the emission probability densities of the HMM. For each user 100 days of calling activity were sampled.

The results with real data are shown in the Figure 1 and with artificial data in the Figure 2. On both figures, the gross profit of the network operator is plotted as

a function of assumed priors. It is relative to the profit from classifying all subscribers as normal and is rescaled by the difference between perfect detection and no detection (one means profit under perfect detection and zero under no detection). The dash dotted line corresponds to the cost-neutral case, the dotted line to the fixed cost classifier, and the solid line to the input-dependent cost model. Experiments were repeated for the following fraud priors: 0.0001, 0.005, 0.001, 0.005, 0.01 (true prior in the artificial data), 0.05, 0.1, 0.133 (true prior in the real data), and 0.5. The values of the assumed priors are plotted on a logarithmic scale, the vertical lines mark the true priors in the data.

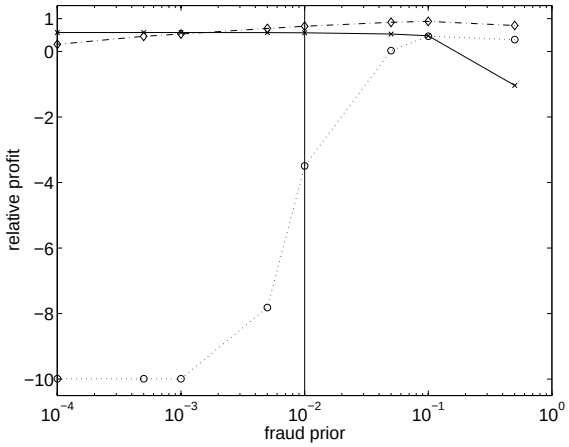

Figure 2: The results from the artificial data experiment.

4 Discussion

In order to make the experiment more realistic, parameters of the cost models were not optimized for the data sets, rather they were educated guesses. In the case of fixed misclassification cost, the relationship between false negative and false positive costs was chosen to be the inverse of the assumed prior to compensate for the imbalanced priors. This seems, according to the experiments, to make this classifier overly sensitive when the assumed prior is very small. In the artificial data experiment, the classifier was still erratic even when the prior was correct. In the real data experiment, when the assumed fraud prior approached the correct value, performance of the fixed cost classifier improved.

The input dependent misclassification cost approach enables incorporating the real costs from the network operator's point of view. Parameters used with this approach were chosen to be the same as the stated profit model used in evaluation.

Those parameters should be readily available to the network operator from the history data. The input dependent misclassification cost approach performed well, although the assumed priors were far away from the ones present in the data. This supports the applicability of this method in real-world problems, since true priors are often unknown and difficult to estimate.

When the model was accurate, as in the artificial data experiment, and the assumed fraud prior was close to the real one, the performance of the cost-neutral classifier was better than others. A similar superiority is also reported in [7]. In the real data experiment, this method also delivered good performance, but it degraded gracefully as the assumed priors moved further away from the correct one.

It is interesting to relate the three methods with the help of Receiver Operating Characteristic (ROC) curves. ROC curve is a function that visualizes the trade-off between false alarms and detection performance for different decision functions [8,9]. In the case of equal costs, the cost can be calculated easily as the sum of errors. Calculating the expected costs with fixed misclassification costs corresponds to a linear mapping from the posteriors by the coefficients in the cost matrix. In the new approach, this mapping is additionally parameterized by the data, enabling the costs to vary from one case to another.

5 Summary

A cost model for input dependent misclassification costs was presented. Experiments were performed in fraud detection within telecommunications domain, where calling behavior was modeled using a simple hidden Markov model. Cost-neutral, fixed misclassification cost and input dependent misclassification cost approaches were used in detecting fraud. The experiments were performed with both real and simulated data and a comparison was made in terms of profit.

The input dependent misclassification cost approach performed favorably especially in the practical problem exemplified by our real data, despite a simplified model and inaccurate priors. Problems in data mining are often characterized by these properties, making this novel method attractive for cost-sensitive classification under the assumption that the input dependent cost model is easily formulated.

Acknowledgments

The first author was funded by Siemens AG. He wishes to thank professor Olli Simula for supervising the graduate studies and also Suvi and Risto for continued support. The work was partially supported by Bundesministerium für Bildung und Forschung, grant number 01IB802A9.

Appendix

In the hidden Markov model, the parameters of the exponential distributions used for modeling the summed length of the of calls per day were for the normal state $\lambda_{s_1} = 3.2$ and for the fraudulent state $\lambda_{s_2} = 10.6$. The initial probability of fraud was zero. The transition matrix for the HMM was set to be

$$P(s_t = j | s_{t-1} = i) = (a_{ji}) = \begin{pmatrix} 0.97 & 0.2 \\ 0.03 & 0.8 \end{pmatrix}.$$

The parameters used in the profit calculation for all methods were 1 and 0.999, for the transaction cost and the discounting of future profit, respectively.

References

[1] Peter D. Turney. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence*, 2:369–409, 1995.

[2] Tom Fawcett and Foster Provost. Adaptive fraud detection. *Journal of Data Mining and Knowledge Discovery*, 1(3):291–316, 1997.

[3] Pedro Domingos. Metacost: A general method for making classifiers cost-sensitive. In Kyuseok Shim and David Madigan, editors, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 155–164, August 1999.

[4] Kazuo J. Ezawa and Steven W. Norton. Constructing bayesian networks to predict uncollectible telecommunications accounts. *IEEE Expert*, 11(5):45–51, October 1996.

[5] Richard O. Duda and Peter E. Hart. *Pattern Recognition and Scene Analysis*. John Wiley & Sons, 1973.

[6] Alan B. Poritz. Hidden markov models: A guided tour. In *Proceedings of the IEEE International conference of Acoustics, Speech and Signal Processing (ICASSP'88)*, pages 7–13, 1988.

[7] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2/3):103–130, November/December 1997.

[8] D.M. Green and J.A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, New York, 1966.

[9] J.P. Egan. *Signal Detection Theory and ROC Analysis*. New York: Academic Press, 1975.