Cost-Sensitive Extensions for Global Model Trees: Application in Loan Charge-Off Forecasting

Marcin Czajkowski[1], Monika Czerwonka[2], and Marek Kretowski[1]

[1] Faculty of Computer Science, Bialystok University of Technology, Wiejska 45a, 15-351 Bialystok, Poland
{m.czajkowski,m.kretowski}@pb.edu.pl
[2] Collegium of Management and Finance, Warsaw School of Economics, Al. Niepodleglosci 162, 02-554 Warsaw, Poland
monika.czerwonka@sgh.waw.pl

Abstract. Most of regression learning methods aim to reduce various metrics of prediction errors. However, in many real-life applications it is prediction cost, which should be minimized as the under-prediction and over-prediction errors have different consequences. In this paper, we show how to extend the evolutionary algorithm (EA) for global induction of model trees to achieve a cost-sensitive learner. We propose a new fitness function which allows minimization of the average misprediction cost and two specialized memetic operators that search for cost-sensitive regression models in the tree leaves. Experimental validation was performed with bank loan charge-off forecasting data which has asymmetric costs. Results show that Global Model Trees with the proposed extensions are able to effectively induce cost-sensitive model trees with average misprediction cost significantly lower than in popular post-hoc tuning methods.

Keywords: cost-sensitive regression, asymmetric costs, evolutionary algorithms, model trees, loan charge-off forecasting.

1 Introduction

In the vast number of contemporary systems, information including business, research and medical issues is collected and processed. In real-life data mining problems, the traditional minimization of prediction errors may not be the most adequate scenario. For example, in medical domain misclassifying an ill patient as a healthy one is usually much more harmful than treating a healthy patient as an ill one and sending him for additional examinations. In finance, investors tend to sell winning stocks more readily than losing stocks in the sense that they realize gains relatively more frequently than losses. The sadness that one experiences in losing the money appears to be greater than the pleasure of gaining the same amount of money. This strong loss aversion was explained and described in the prospect theory by Kahneman and Tversky [14] and applied to finance practice by Shefrin and Statman [25].

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_30,

In this paper, we want to tackle the cost-sensitive regression methods. We focus on extending the existing *EA* for model tree induction to handle data with asymmetric costs.

1.1 Background

The decision trees [22] are one of the most widely used prediction techniques. Ease of application, fast operation and what may be the most important, effectiveness of decision trees, makes them powerful and popular tool [15]. Regression and model trees [13] may be considered as a variant of decision trees, designed to approximate real-valued functions instead of being used for classification tasks. The main difference between regression tree and model tree is that, in the latter, constant value in the terminal node is replaced by a regression plane. Each leaf of the model tree may hold a linear (or nonlinear) model whose output is the final prediction.

Problem of learning an optimal decision tree is known to be NP-complete. Consequently, classical decision-tree learning algorithms are built with a greedy top-down approach [21] which usually leads to suboptimal solutions. Recently, application of *EA*s [18] to the problem of decision tree induction [2] become increasingly popular alternative. Instead of local search, *EA* performs a global search in the space of candidate solutions. Trees induced with *EA* are usually significantly smaller in comparison to greedy approaches and highly competitive in terms of prediction accuracy [17,7]. On the other hand, the induction of global regression and model trees is much slower [8]. One of the possible solutions to speed up evolutionary approach is a combination of *EA*s with local search techniques, which is known as Memetic Algorithms [12].

Cost-sensitive prediction is the term which encompasses all types of learning where cost is considered [28,10] e.g., costs of tests (attributes), costs of instances, costs of errors. In this paper, we only focus on asymmetric costs, which are associated with different types of prediction errors.

The vast majority of data mining algorithms is applied only to the classification problems [27] while cost-sensitive regression is not really studied outside of statistic field [3]. In induction of cost-sensitive classification trees, three techniques are popular:

- convert classical decision tree into cost-sensitive one, mainly by changing the splitting criteria and/or adopting pruning techniques for incorporating misclassification costs (e.g. [4]);
- application of *EA*s that induce cost-sensitive trees [16];
- application of universal methods like: cost instance-weighting [26] or post-hoc tuning solutions e.g. MetaCost [9].

One of the earliest studies of asymmetric costs in regression was performed by Varian [30]. Author propose *LinEx* loss function which is approximately linear on one side and exponential on the other side as an alternative to popular least squared procedures. Application of different loss functions was later extended

[5] to *LinLin* (asymmetric linear) and QuadQuad (asymmetric quadratic) loss functions. In data mining literature there are only few propositions to handle asymmetric costs e.g. in [6] authors propose a modified back-propagation neutral network that applies *LinLin* cost function.

Recently, post-hoc tuning methods for regression, analogous to ones in cost-sensitive classification, were proposed [3]. Solutions minimize average misprediction cost under an asymmetric cost structure for regular regression models post-hoc by adjusting the prediction by a certain amount. In it's extension [31], application of polynomial functions as model adjustment is proposed to improve the cost-sensitive prediction.

1.2 Motivation

Due, to the lack of cost-sensitive regression solutions in data mining literature, one of the good alternatives are the post-hoc tuning methods [3,31]. However, limitations of such algorithms are obvious as the tuning procedure cannot incorporate cost functions during model learning. In addition, when understanding and interpretation of generated decisions/rules is crucial, such technique cannot be applied.

In this paper, we want to show how to extend existing evolutionary induced model trees to successfully predict under asymmetric losses. In case of evolutionary induced model trees, simple modification of the fitness function, alike for classification trees [17] is not enough, as the linear (or non-linear) models in the leaves are usually not evolved but constructed using standard regression techniques [1,7]. Extensions must also affect the search of cost-sensitive models in the leaves. Full search of regression models is usually difficult for real-life, large datasets due to the huge additional solution space to cover. Therefore, in this paper, we propose two memetic operators that can, together with appropriate fitness function, efficiently convert cost-neutral model trees into cost-sensitive ones.

2 Cost-Sensitive Extensions for Evolutionary Induced Model Trees

In this section we present a combination of evolutionary approaches with local search techniques to achieve a cost-sensitive learner. At first, we briefly describe evolutionary evolved model tree called Global Model Tree (*GMT*) [7]. This evolutionary induced model tree will serve as an example to illustrate the proposed extensions and the fitness function to handle data with asymmetric costs.

2.1 Global Model Tree

GMT follows a typical framework of evolutionary algorithms [18] with an unstructured population and a generational selection. Model trees are represented in their actual form as typical univariate trees. Each test in a non-terminal node

concerns only one attribute (nominal or continuous valued). At each leaf a multivariate linear model is constructed using standard regression technique [20] with instances and attributes associated with that node.

Initial individuals are created by applying the classical top-down algorithm [21]. Ranking linear selection [18] is used as a selection mechanism. Additionally, in each iteration a single individual with the highest value of fitness function in current population in copied to the next one *(elitist strategy)*. Several variants of cross-over and mutations were proposed [7,8] that involve:

- exchanging tests, nodes, subtrees and branches between the nodes of two individuals;
- modifications in the tree structure (pruning the internal nodes and expanding the leaves);
- changing tests in internal nodes and extending, simplifying, changing linear regression models in the leaves.

The Bayesian information criterion (BIC) [23] is used as a fitness function and its formula is given by:

$$Fit_{BIC}(T) = -2 * ln(L(T)) + ln(n) * k(T), \tag{1}$$

where $L(T)$ is maximum of likelihood function of the tree T, $k(T)$ is the number of model parameters and n is the number of observations. The log(likelihood) function $L(T)$ is typical for regression models [11] and can be expressed as:

$$ln(L(T)) = -0.5n * [ln(2\pi) + ln(SS_e(T)/n) + 1], \tag{2}$$

where $SS_e(T)$ is the sum of squared residuals on the training data of the tree T. The term $k(T)$ can also be viewed as a penalty for over-parametrization and has to include not only the tree size (calculated as the number of internal nodes) but also the number of attributes that build models in the leaves.

2.2 Cost-Sensitive Extensions

Extending regular regression models to be cost-sensitive requires several steps. At first, appropriate measurement must be defined for assessing the performance of solutions. In our work, we use the average misprediction cost proposed in [3].

Let the dependent variable y be predicted based on a vector of independent variables x. A regression method learns a prediction model, $f : x \rightarrow y$ from n training instances. If the function $C(e)$ characterize the cost of a prediction error e then average misprediction cost denoted as Amc can be defined as:

$$Amc = \frac{1}{n} \sum_{1}^{n} C(f(x_i) - y_i). \tag{3}$$

Next, to find cost-sensitive regression models in the tree leaves, we propose BIC extension as fitness function and two local search components that are built into the mutation-like operator.

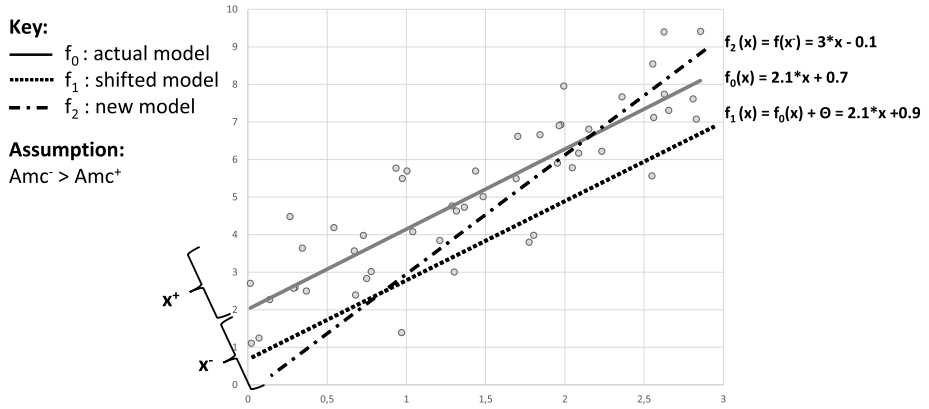

Fig. 1. An example of simple linear regression model $f_0(x)$ changed by cost-sensitive extensions - shift: $f_1(x)$ and new model: $f_2(x)$

Fitness Function. We propose a cost-sensitive BIC to work as a fitness function. We have replaced the squared error loss $SS_e(T)$ from Equation 2 with the average misprediction cost. To remain balance between complexity term $k(T)$ and the cost of the tree, we performed additional experimental research to determine the appropriate value of penalty term, which is now equal $(Q(T) + M(T))$ where $Q(T)$ is the number of internal nodes in model tree T and $M(T)$ is the sum of all attributes in the linear models in the leaves.

Shift Regression Model. The idea of our first mutation variant is similar to the one for cost-sensitive post-hoc tuning method [3]. With the user defined probability, regression model in the leaf is adjusted by a certain amount denoted as θ. Let x^+ represents instances that are over-predicted and x^- instances under-predicted by an actual regression model in the leaf. The costs for over-prediction and under-prediction are equal C^+ and C^-, respectively.

We calculate average misprediction cost separately for x^+ and x^-, denoted as Amc^+ and Amc^- and define the shift θ as:

$$\theta = \begin{cases} -\frac{Amc^+}{C^+} * \delta, \text{ if } Amc^+ > Amc^- \\ \\ \frac{Amc^-}{C^-} * \delta, \text{ if } Amc^+ < Amc^- \end{cases}, \tag{4}$$

where δ is equal:

$$\delta = \frac{Amc^+ - Amc^-}{Amc^+ + Amc^-} rand(0, 1). \tag{5}$$

Main role the parameter δ is to reduce impact of adjustment when Amc on both sides of regression model is similar. Multiplication with a random value from 0 to 1 (denoted as $rand(0, 1)$) extends the number of possible values of θ. Finally,

the actual regression model in the leaf is updated by adding the calculated adjustment:

$$f_{new}(x) = f(x) + \theta. \tag{6}$$

It is illustrated in Figure 1 where actual regression model $f_0(x)$ is replaced by the shifted one $f_1(x)$.

New Cost-Sensitive Model. Second variant of mutation replaces actual regression model with a new one that is built on the subset of instances. If, for the actual model in the leaf, the $Amc^+ > Amc^-$ then new cost-neutral regression model is calculated only for over-predicted instances (x^+), otherwise only for under-predicted (x^-). Next, the actual regression model is replaced by the new one:

$$f_{new}(x) = \begin{cases} f(x^+), \text{ if } Amc^+ > Amc^- \\ \\ f(x^-), \text{ if } Amc^+ < Amc^- \end{cases}. \tag{7}$$

In contrast to the first extension, this technique allows finding a completely new model that can decrease Amc for the leaf. Figure 1 illustrates how actual regression model $f_0(x)$ is replaced by the new one denoted as $f_2(x)$, calculated for the x^-.

3 Experiments

We have modified cost-neutral GMT algorithm to show, how proposed extensions handle data with asymmetric costs. In this section we show the performance of $CS-GMT$ (GMT with applied cost-sensitive extensions) on loan charge-off forecasting data. Thanks to the source code of cost-sensitive tuning method and its extensions received from authors [3,31] we are able to compare $CS-GMT$ with post-hoc tuning methods.

3.1 Datasets and Setup

In the paper we used loan charge-off forecasting data from Wharton Research Data Services ($WRDS$, http://wrds-web.wharton.upenn.edu). This data is characterized by asymmetric costs on misprediction errors, because under-prediction of loan charge-off is more costly than over-prediction. If the bank over-predicts its future loan charge-off, the worst what could happen is the reduction of bank's income because there will maintain some extra funds in the loan-loss reserves. The under-prediction means that the bank did not prepare sufficient provisions for its loan losses and has not enough reserves which can cause regulatory problems and significant downturn of its credit rating which is much more dangerous to the bank.

We used the same settings to prepare and test data as in [31], however, more recent data were used. In the experiments, 28 quarters from period $2004-2010$ were used with 14 variables related to bank current financial data (described

and listed in [3,31]), including loan charge-off, in a particular quarter as the independent variable. The dependent variable is the loan charge-off in the following quarter so the bank can use all useful information while predicting the next quarter loan charge-off.

We generated 27 datasets from 28 quarters because from the last quarter of 2010 only loan charge-off value were used as 2011 data were not available in $WRDS$ yet. For each dataset, prediction model was trained on one quarter and tested on the next one and so on. Therefore, there were 26 training datasets (third quarter of 2010 was used only for testing) and 26 independent testing datasets (first quarter of 2004 was used only for training). In addition, observations with missing values were removed and, to reduce the extent of skewness, the natural logarithm transformation was performed. Average number of instances in each quarter equals 7695 (minimum: 6992 and maximum: 8315). Following [31], we used $LinLin$ cost function and examined cost ratios for under-prediction to over-prediction as follows: $10:1$, $20:1$, $50:1$ and $100:1$. The same three base regression models: standard least-squares linear regression (LR), $M5$ model tree [29] and back-propagation neutral network (NN)[24] were post-hoc tuned for the comparison purpose to $CS-GMT$. Original settings for all tuned methods and $CS-GMT$ solution were applied through all experiments.

3.2 Results

Table 1 summarizes the results of the Amc for three base regression methods tuned by the algorithms described in [3,31] and proposed $CS-GMT$ solution. Each reported quantity is an average value over 26 independent testing datasets (over 200 000 tested instances). The $NONE$ column refers to the results without tuning or cost-sensitive extensions, BSZ refers to the tuning method proposed by Bansal et al. [3] and $LINEAR$ is a linear extension of BSZ algorithm by Zhao et al. [31]. Finally, last column shows the results of $CS-GMT$: proposed cost-sensitive extensions denoted as CS *extensions* applied to GMT.

Results enclosed in Table 1 show that the both post-hoc tuning methods improves the performance of regression model. The extension of BSZ called $LINEAR$, like it was shown in the paper [31], is significantly better than its predecessor. When only post-hoc tuned algorithms are considered, we can observe that the best performance is achieved by NN. However, when we focus on the last column, we see that Amc can be decreased even more. The $CS-GMT$ solution outperforms all tuned base regression models under every cost ratio. Wilcoxon signed rank test for $CS-GMT$ and linearly tuned NN under every cost ratio showed that the differences of Amc between both algorithms are statistically significant ($P\ value < 0.0001$). There is also a significant difference between linearly tuned GMT and $CS-GMT$ which suggests that there is a still significant space for improvement for tuned methods.

The cost reduction between the best out of three linearly tuned algorithms (NN) and $CS-GMT$ is in the range of 7.7% to 9.4% which may be seen by some as not very impressive. However, we must remember that the cost values are on a natural log scale as the values of dependent variable loan charge-off

Table 1. Average misprediction costs for post-hoc tuned base regression algorithms and cost-sensitive extensions for Global Model Tree

Algorithm	Cost ratio	NONE	BSZ	LINEAR	CS extensions
LR	10	7.41	3.78	3.81	-
M5	10	7.29	4.16	3.88	-
NN	10	8.16	3.69	3.57	-
GMT	10	7.07	3.81	3.65	3.29
LR	20	14.06	4.84	4.42	-
M5	20	13.78	5.47	4.86	-
NN	20	15.60	4.62	4.26	-
GMT	20	13.66	5.07	4.27	3.85
LR	50	34.02	6.24	5.23	-
M5	50	33.23	6.03	6.44	-
NN	50	37.92	5.69	5.09	-
GMT	50	32.96	6.40	6.11	4.66
LR	100	67.27	7.06	5.85	-
M5	100	65.66	7.24	7.94	-
NN	100	75.12	6.50	5.80	-
GMT	100	64.15	7.03	5.98	5.27

were transformed by the natural logarithm. Therefore, the real cost reduction on the original scale is in range 24.2% to 40.7% and therefore, can be attractive in bank loan charge-off forecasting problem.

Application of proposed cost-sensitive extensions does not only significantly reduce *Amc*. The important benefit of proposed extensions, in context of GMT is that the whole tree: test in internal nodes and models in the leaves, fits to the analyzed, cost-sensitive problem. Therefore, decisions from $CS-GMT$ are much easier to interpret. Identifying patterns and finding explanations for predictions may be difficult for tuned regression models because all rules obtained by the phase of learning were cost-neutral.

4 Conclusion and Future Works

In this paper, we propose extensions for evolutionary induced model trees to achieve cost-sensitive learners. We present a cost-sensitive BIC to work as the fitness function and allows the algorithm to minimize the average misprediction cost. Two specialized memetic operators that search for cost-sensitive regression models in the tree leaves are also proposed. Those local optimizations of the regression models are simple, complementary and easy to apply.

Experiments performed of 27 real-life datasets show that there is a significant difference between post-hoc tuned methods and solutions that explicitly incorporate cost functions during model building - like proposed $CS-GMT$. The real, average cost reduction between the best out of 4 tuned algorithms (linearly tuned NN) and proposed $CS-GMT$ is over one third. In addition, as generated

decisions and models from $CS - GMT$ take into account costs during learning phase, they can be used to learn and understand underlying processes from the data.

There are a number of promising directions for future research. In particular, we should consider testing different cost functions and handling different types of costs like e.g. cost of attributes. Application of the proposed extensions to other EA solutions that construct models with standard regression techniques and testing other forecasting problems requires further extensive research, which we leave for the future.

Acknowledgments. This work was supported by the grant W/WI/1/2013 from Bialystok University of Technology. The authors thank to Huimin Zhao, Atish P. Sinha and Gaurav Bansel for providing us with implementation of their tuning methods.

References

1. Barros, R.C., Ruiz, D.D., Basgalupp, M.: Evolutionary model trees for handling continuous classes in machine learning. Information Sciences 181, 954–971 (2011)
2. Barros, R.C., Basgalupp, M.P., Carvalho, A.C., Freitas, A.A.: A Survey of Evolutionary Algorithms for Decision-Tree Induction. IEEE Transactions on Systems Man and Cybernetics, Part C 42(3), 291–312 (2012)
3. Bansal, G., Sinha, A.P., Zhao, H.: Tuning data mining methods for cost-sensitive regression: a study in loan charge-off forecasting. Journal of Management Information Systems 25(3), 317–338 (2008)
4. Bradford, J., Kunz, C., Kohavi, R., Brunk, C., Brodley, C.E.: Pruning decision trees with misclassification costs. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 131–136. Springer, Heidelberg (1998)
5. Cain, M., Janssen, C.: Real estate price prediction under asymmetric loss. Annals of the Institute of Statistical Mathematics 47(3), 401–414 (1995)
6. Crone, S.F., Lessmann, S., Stahlbock, R.: Utility based data mining for time series analysis: Cost-sensitive learning for neural network predictors. In: Proc. of 1st UDBM, Chicago, IL, pp. 59–68 (2005)
7. Czajkowski, M., Kretowski, M.: An Evolutionary Algorithm for Global Induction of Regression Trees with Multivariate Linear Models. In: Kryszkiewicz, M., Rybinski, H., Skowron, A., Raś, Z.W. (eds.) ISMIS 2011. LNCS, vol. 6804, pp. 230–239. Springer, Heidelberg (2011)
8. Czajkowski, M., Kretowski, M.: Does Memetic Approach Improve Global Induction of Regression and Model Trees? In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) SIDE 2012 and EC 2012. LNCS, vol. 7269, pp. 174–181. Springer, Heidelberg (2012)
9. Domingos, P.: MetaCost: A general method for making classifiers cost-sensitive. In: Proc. of KDD 1999, pp. 155–164. ACM Press (1999)
10. Elkan, C.: The Foundations of Cost-Sensitive Learning. In: Proc. of IJCAI, pp. 973–978 (2001)
11. Gagne, P., Dayton, C.M.: Best Regression Model Using Information Criteria. Journal of Modern Applied Statistical Methods 1, 479–488 (2002)

12. Gendreau, M., Potvin, J.Y.: Handbook of Metaheuristics. International Series in Operations Research & Management Science, vol. 146 (2010)
13. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd edn. Springer (2009)
14. Kahneman, D., Tversky, A.: Prospect Theory: An Analysis of Decisions under Risk. Econometrica 47(2), 263–292 (1979)
15. Kotsiantis, S.B.: Decision trees: a recent overview. Artificial Intelligence Review, 1–23 (2011)
16. Krętowski, M., Grześ, M.: Evolutionary Induction of Cost-Sensitive Decision Trees. In: Esposito, F., Raś, Z.W., Malerba, D., Semeraro, G. (eds.) ISMIS 2006. LNCS (LNAI), vol. 4203, pp. 121–126. Springer, Heidelberg (2006)
17. Kretowski, M., Grześ, M.: Evolutionary Induction of Mixed Decision Trees. International Journal of Data Warehousing and Mining 3(4), 68–82 (2007)
18. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs, 3rd edn. Springer (1996)
19. Murthy, S.: Automatic construction of decision trees from data: A multi-disciplinary survey. Data Mining and Knowledge Discovery 2, 345–389 (1998)
20. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: Numerical Recipes in C. Cambridge University Press (1988)
21. Rokach, L., Maimon, O.Z.: Top-down induction of decision trees classifiers - A survey. IEEE Transactions on Systems Man and Cybernetics, Part C 35(4), 476–487 (2005)
22. Rokach, L., Maimon, O.Z.: Data mining with decision trees: theory and application. Machine Perception Arfitical Intelligence 69 (2008)
23. Schwarz, G.: Estimating the Dimension of a Model. The Annals of Statistics 6, 461–464 (1978)
24. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. In: Parallel Distributed Processing, pp. 318–362. MIT Press, Cambridge (1986)
25. Shefrin, H., Statman, M.: The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence. Journal of Finance 40, 777–790 (1985)
26. Ting, K.: An instance-weighting method to induce cost-sensitive trees. IEEE Transactions on Knowledge and Data Engineering 14(3), 659–665 (2002)
27. Torgo, L., Ribeiro, R.: Utility-based regression. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 597–604. Springer, Heidelberg (2007)
28. Turney, P.: Types of cost in inductive concept learning. In: Proc. of ICML 2000 Workshop on Cost-Sensitive Learning, Stanford, CA (2000)
29. Quinlan, J.: Learning with Continuous Classes. In: Proc. of AI 1992, pp. 343–348. World Scientific (1992)
30. Varian, H.R.: A Bayesian Approach to Real Estate Assessment. In: Fienberg, S.E., Zellner, A. (eds.) Studies in Bayesian Econometrics and Statistics: In honor of L.J. Savage, North-Holland, Amsterdam, pp. 195–208 (1974)
31. Zhao, H., Sinha, A.P., Bansal, G.: An extended tuning method for cost-sensitive regression and forecasting. Decision Support Systems 51, 372–383 (2011)