Cost-Sensitive Bayesian Network Learning using Sampling

Eman Nashnush[1] , Sunil Vadera[1]

[1]The school of Computing, Science and Engineering, Salford university , Manchester, UK.
E.Nashnush1@edu.salford.ac.uk , S.Vadera@salford.ac.uk

Abstract. A significant advance in recent years has been the development of cost-sensitive decision tree learners, recognising that real world classification problems need to take account of costs of misclassification and not just focus on accuracy. The literature contains well over 50 cost-sensitive decision tree induction algorithms, each with varying performance profiles. Obtaining good Bayesian networks can be challenging and hence several algorithms have been proposed for learning their structure and parameters from data. However, most of these algorithms focus on learning Bayesian networks that aim to maximise the accuracy of classifications. Hence an obvious question that arises is whether it is possible to develop cost-sensitive Bayesian networks and whether they would perform better than cost-sensitive decision trees for minimising classification cost? This paper explores this question by developing a new Bayesian network learning algorithm based on changing the data distribution to reflect the costs of misclassification. The proposed method is explored by conducting experiments on over 20 data sets. The results show that this approach produces good results in comparison to more complex cost-sensitive decision tree algorithms.

Keywords: Cost-sensitive classification, Bayesian Learning, Decision Trees.

1 Introduction

Classification is one of the most important methods in data mining; playing an essential role in data analysis and pattern recognition, and requiring the construction of a classifier. The classifier can predict a class label for an unseen instance from a set of attributes. However, the induction of classifiers from the data sets of pre-classified instances is a central problem in machine learning [1]. Therefore, several methods and algorithms have been introduced, such as decision trees, decision graphs, Bayesian networks, neural networks, and decision rules, etc. Over the last decade, graphical models have become one of the most popular tools to structure uncertain knowledge. Furthermore, over the last few years, Bayesian networks have become very popular and have been successfully applied to create consistent probabilistic representations of uncertain knowledge in several fields [2].

Cost-insensitive learning algorithms focus only on accuracy (class label output), and do not take misclassification costs or test costs into consideration. However, the performance of any learning algorithm, in practice, normally has to take the cost of misclassification into account. Hence, in recent years, a significant level of attention has been paid to cost-sensitive learning, including making accuracy-based learners cost-sensitive [3, 4]. Zadrozny *et al*. [6] divide cost-sensitive classifiers into two

categories: the amending approach (changing the classifier to a transparent box) and the sampling approach (using the classifier as a black box). Among all the available cost-sensitive learning algorithms, most of the work has focused on decision tree learning, with very few studies considering the use of Bayesian networks for cost-sensitive classification.

Therefore, this paper aims to explore the use of Bayesian networks (BNs) for cost-sensitive classification. During this paper, a new method known as the Cost-Sensitive Bayesian Network (CS-BN) algorithm, which uses a sampling approach to induce cost-sensitive Bayesian networks, is developed and compared with other, more common approaches such as cost-sensitive decision trees. This paper is organized as follows: in section 2 we will provide a number of definitions and background information on cost-sensitive learning algorithms. Section 3 will introduce some of the previous work on the sampling approach. In section 4, we will present our method for converting the existing BN algorithm into a CS-BN by changing the number of examples to reflect the costs. In section 5 we present the results obtained by carrying out an empirical evaluation on data from the UCI repository. Finally, section 6 will provide a conclusion, along with a summary of the main contribution of this paper.

2 Cost-sensitive learning perspective and overview

A good cost-sensitive classifier should be able to predict the class of an example that leads to the lowest expected cost, where the expectation is computed after applying the classifier by using the expected cost function, as shown in the following equation [6,21]:

$$expected\ cost(x|i) = \sum_{j} C(i,j)P(j|x). \qquad (1)$$

Where $P(j|x)$ represents the probability of an example x being in class j given it is actually of class *i,* and $C(i,j)$ represents the cost of misclassifying an example as class *i* when it is in class *j* [21]. In particular, cost-sensitive algorithms aim to minimize the number of high-cost misclassification errors, thus reducing the total number of misclassification errors. According to Zadrozny *et al.* [6], cost-sensitive classifiers can be divided into two categories: *Black Box* (sampling), and *Transparent Box.* Black box methods use a classifier as a black box, and use resampling methods according to a class weight. On the other hand, transparent box methods use weights to change the classifier learning algorithm directly. Conversely, Sheng and Ling [7] used different terms such as *direct method*, and *wrapper methods,* as shown in Figure1.

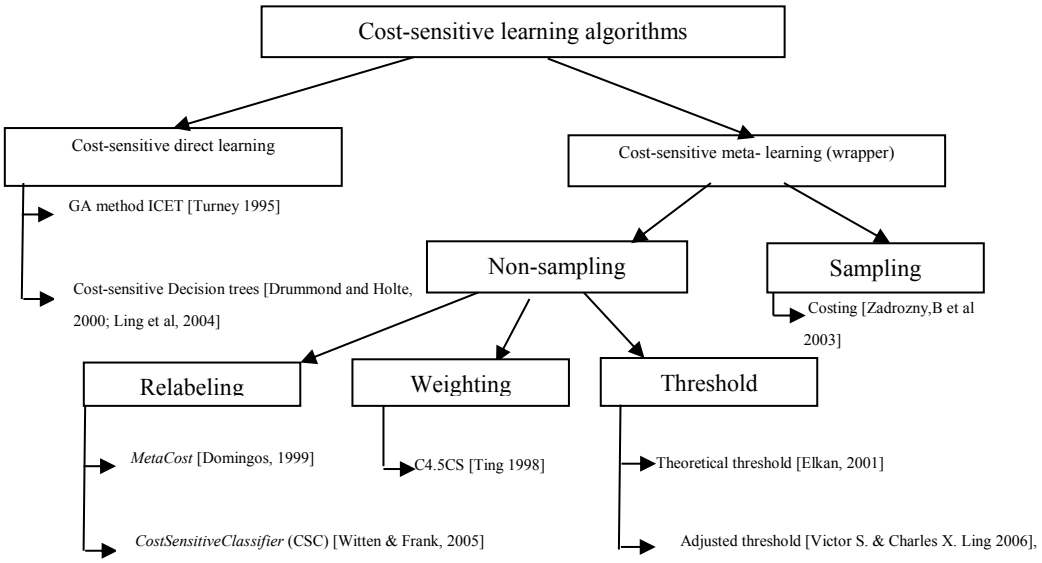

Fig.1. Cost-sensitive learning category (Shend and Ling [7])

As Zadrozny [6] points out, wrapper methods (Black Box), deal with a classifier as a closed box, without changing the classifier behaviour, and can work for any classifier. In contrast, direct methods (Transparent Box), require knowledge of the particular learning algorithm, and can also work on the classifier itself by changing its structure to include the costs.

3 Review of previous work on sampling approach

Most studies regarding cost-sensitive learning have used direct methods or sampling methods, and most have focused on decision tree learning. This section briefly reviews some of these methods. In addition, this section describes different methods of cost-sensitive learning by changing the data distribution to involve costs and solve an unbalanced data distribution problem, where, for example, the number of negative examples is significantly less than the number of positive examples. Several literature reviews show different methods, where some of them amend the number of negative examples (*over-sampling*); some of them change the number of positive examples (*under-sampling*); a few of them use the "*SMOTE*" (Synthetic Minority Over-Sampling Technique) algorithm that tackles the imbalanced problem by generating synthetic minority class examples [8]; and others use a "*Folk Theorem*" [5, 21] that amends the distribution according to the cost of misclassification.

Kubat and Matwin [12] used one side selection by under-sampling the majority class while keeping the original population of the minority class. As Elkan [21] pointed out in 2001, changing the balance of negative and positive training examples will affect classification algorithms. Ling and Li [13] combined over-sampling with under-sampling to measure the performance of a classifier. Domingos [14] introduced the MetaCost algorithm which is based on sampling with labeling and bagging. MetaCost uses the resampling with replacement to create a different sample

size, then estimates each example in the same sample size by voting each example in different samples, where the number of instances in each resamples is smaller than the training size, and then applies an equation (1) to re-label each training example with the optimal class estimation. Finally, it reapplies the classifier again, on the new relabelled training data set [15]. Figure 2 summarises the MetaCost algorithm [15]. Domingos concluded that this algorithm provides goods results on large data sets. In addition, most researchers have dealt with this problem by changing the data distributions to reflect the costs, though most of them utilize a decision tree learner as a base learner, and the reader is referred Lomax and Vadera [4] for a comprehensive survey of cost sensitive decision tree algorithms for details.

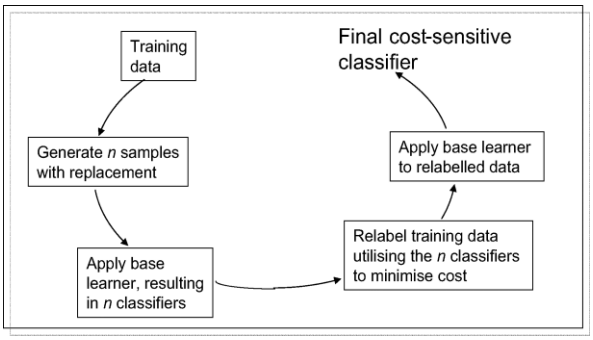

Fig. 2. The MetaCost system [15]

4 New Cost-Sensitive Bayesian Network learning algorithm via the distributed sampling approach

A survey of the literature shows that, to date, there are very few publications regarding cost sensitive Bayesian networks (CS-BNs), but plenty on cost-sensitive decision tree learning. This section presents a sampling approach used to develop CS-BNs and presents the use of distributed sampling to take account of misclassification costs and reduce the number of errors. Thus, the compelling question, given the different class distributions, is: what is the correct distribution for a learning algorithm?

In response, it has been observed that naturally-occurring distributions are not always the optimal distribution [8]. In our experiments, we used the sampling (Black Box) method, because this method can also be used to address the imbalanced data problem and can be applied to any learning algorithm. In our study, we used Folk Theorem to change the data distribution. This approach has previously been introduced by Zadrozny *et al.* [5]. This theorem draws a new distribution from the old distribution, according to cost proportions, to change the data distribution and obtain optimal cost-minimization from the original distribution. This theorem is only theoretically motivated, and does not require any probability density estimation. Thus, we have used this theorem on the BN classifier, which has not been used before in this classifier.

4.1 Use of the Folk Theorem for CS-BNs

This method can be applied to any cost-insensitive classification algorithm to form a cost-sensitive classification algorithm. This method can be conducted by reweighing the instances from the training example and then using that weight on the classification algorithm. The Folk Theorem is used to change the data distribution to reflect the costs. Zadroznyet *al.*, [5] stated that "*if the new examples are drawn from the old distribution, then optimal error rate classifiers for the new distributions are optimal cost minimizes for data drawn from original distribution.*" This is shown in the following equation (2) [5]:

$$D'(\mathrm{x},\mathrm{y},\mathrm{c}) = \frac{\mathrm{C}}{\mathrm{E}_{\mathrm{x,y,c\sim D[c]}}}\mathrm{D}(\mathrm{x},\mathrm{y},\mathrm{c}). \qquad (2)$$

Where, the new distribution D' = factor * Old distribution D; x is instance; y is the class label; and C is the cost according to misclassified instance x. Technically, the optimal error rate classifiers from D' are the optimal cost minimizers from the data, which have been drawn from D. This theorem creates new distribution from the old distribution by multiplying old distribution with a factor proportional to the relative cost of each example, and the new distribution will be adapted with that cost. Therefore, this method enables the classifier to obtain the expected cost minimization from the original distribution and, in the worst case scenario; this method can be guaranteed a classifier to provide a good approximate cost minimization for any new sample.

However, there are different types of BNs, as well as methods for learning them. Given their efficiency compared to full networks, we used a search algorithm to construct *Tree Augmented Naive Bayes Networks* (TANs), along with *Minimum Description Length* (MDL), which was introduced by Fayyad and Irani [19] to calculate the score of information between links in a tree.

In our experiments, we attempted to change the proportions of instances (samples) in each class label, according to its cost, by using the above Folk Theorem [5]. In the current experiment, we used a constant cost of 1:4, where we assigned the common majority class cost to 1 and other, minority class cost to 4. The following steps were conducted with the CS-BN by using a sampling approach:

- **Splitting**: Data are split into a training set and testing set. The training set uses 75% of the original data, while the testing set uses 25% of the original data.
- **Cost proportion**: According to cost proportions, the new data distribution should be calculated as being equal to these proportions. For instance, if the cost of wrongly classifying a sick patient as healthy is £20 and the cost of misclassifying a healthy patient as sick is £2, then the cost proportion of the sick class will be 20/22=0.90. In our experiments we used cost proportion by assigning rare class cost to 4 and common class cost to 1. Thus, the cost proportion in our algorithm would be 0.8 and 0.2 respectively, based on equation (3):

$$CostProportionofclass_i = \frac{Cost_i}{\sum_{j=0}^{k} Cost_j}. \qquad (3)$$

Where i and j is the class index and k is the number of classes.

- **Changing proportion**: This involves changing the training data distributions according to the cost ratio of each class. For example, when the costs are 1:4, the new proportions on the training set for each class will be 20% and 80% respectively.

There are different methods that can be used to achieve sampling. During our research, we used two methods, as discussed in section 3 of this paper. These methods were *under-sampling* and *over-sampling*. Obviously, where the new proportion was less than the original proportion, we used under-sampling (without replacement) to delete some of the examples in the frequent class. On the other hand, if the new proportion was greater than the original proportion, we used over-sampling (with replacement) by making a random generation of new instances which belonged to the rare class, and increasing the number of examples. As a result, the training data required further resampling according to their costs. Finally, we used the original BN classifier on the training data, followed by using the testing set with the original distribution (without changing any instances) to evaluate the training model.
However, Figure 3 presents the pseudo-code of our method (i.e. CS-BN with sampling approach):

CS-BN via sampling approach:

1. Divide dataset into 75% of instances for training, and 25% for testing. With the same class distributions.
2. Changing the data distribution according to the cost proportions in each class, $CostProportion = \frac{costi}{\sum_{i,j} costi,j}$
3. Using Bayesian network algorithm (TAN to learn structure).
4. Evaluating the model on the original test set distribution.

Fig. 3. Cost-Sensitive Bayesian Network Algorithm by sampling

4.2 Experiment

Our experiment demonstrates how changing the distribution of data will affect the performance and cost of a Bayesian classifier. We experimented with 24 data sets from the UCI repository [20]. To evaluate the performance of our proposed method, we used the original testing distribution. An evaluation was carried out in order to compare CS-BN with existing algorithms implemented in WEKA: (i) Original Bayes Net (that implemented by TAN) (Friedman *et al.* [1], Version 8); (ii) Decision Tree Algorithm J48 (which is their implementation of C4.5, Version 8); and (iii) MetaCost with J48 as the base classifier [14](iv) Naive Bayes. Table 1 presents the results of the CS-BN algorithm via changing distributions (Black Box), and the original BN algorithm. It also shows the comparison between the original Bayes Net (TAN), existing algorithm (decision tree J48), MetaCost with j48 classifier, and NB. The proposed algorithm produced lower costs for cost matrix 1 and 4 on most of the data set. In our experiment, we noticed that number of False Negative (rare) with our Black Box method was less than number of False Negative (rare) of the existing BN algorithm; thus, the total cost will be around 6121 units, and we reduce FN in all datasets.

	CS-BN using Sampling				Original Bayes Network				Metacost j48 CSC+J48		Original Decision tree J4.8		Original Naïve Bayes NB	
Dataset	FN	FP	Accuracy	Cost	FN	FP	Accuracy	Cost	Accuracy	Cost	Accuracy	Cost	Accuracy	Cost
Adult	**109**	3420	71.01	**3856**	994	704	86.05	4680	81.45	3812	86.014	5300	83.28	6353
Horse-colic	**1**	31	65.59	**35**	26	7	64.51	111	63.44	40	66.66	85	64.51	45
Austra	**6**	29	79.76	**53**	11	13	86.12	57	84.97	41	84.39	84	75.14	127
Bank	**2**	66	55.26	**74**	18	19	75.65	91	91.44	34	91.44	34	78.94	95
Breast	**0**	3	98.31	**3**	1	3	97.75	7	94.94	21	94.94	21	96.06	13
Bupa liver	**0**	51	42.04	**51**	37	0	57.95	148	56.81	56	63.63	71	60.22	53
Crx	**6**	25	82.08	**49**	13	10	86.70	62	83.23	47	87.28	61	75.72	147
Diabetes	**6**	69	61.13	**93**	19	24	77.72	100	73.05	118	74.09	134	78.75	113
German	**2**	106	57.14	**114**	43	29	71.42	201	70.63	143	71.82	218	74.20	179
Gymexamg	**0**	439	40.09	**439**	189	0	69.90	756	36.62	452	67.83	742	67.35	736
Heart	**3**	14	75.36	**26**	6	3	86.956	27	79.71	44	79.71	44	79.71	35
Hepati	**0**	8	80.48	8	1	2	92.68	6	80.48	11	70.73	27	82.92	10
Horse	**8**	13	77.65	**45**	13	3	82.97	55	77.65	60	79.78	64	79.78	55
Hoslem	**0**	0	100	**0**	0	0	100	0	100	0	100	0	85.71	16
Hypo	**2**	10	98.86	**18**	3	6	98.86	18	98.86	27	98.99	26	97.98	46
Iono	**6**	7	85.55	**31**	7	3	88.88	31	86.66	30	86.66	30	76.66	42
labor	**0**	1	93.33	**1**	0	1	93.33	1	80	9	80	9	93.33	1
Pima	**7**	62	64.24	**90**	26	19	76.68	123	72.53	101	72.53	101	80.82	100
Sonar	**11**	7	66.66	**51**	15	5	62.96	65	61.11	63	61.11	63	59.25	61
Spambase	**32**	57	92.20	**185**	45	32	93.25	212	91.76	241	92.03	238	78.63	298
Supermarket	**0**	735	36.25	**735**	418	0	63.74	1672	36.25	735	63.74	1672	63.74	1672
Tic-tac	**6**	112	51.23	**136**	36	27	73.96	171	81.81	89	84.29	92	66.11	229
Unbalanced	**3**	0	98.59	**12**	3	0	98.59	12	98.59	12	98.59	12	92.05	23
Vote	**2**	6	91.81	**14**	3	2	95.45	14	93.63	10	95.45	11	90	14
Weather	**0**	2	40	**2**	0	3	40	3	60	2	60	2	100	0
Total cost				6121				8623		6198		9141		10463

Table 1.Comparison between CS-BN via changing the distributions and existing algorithms

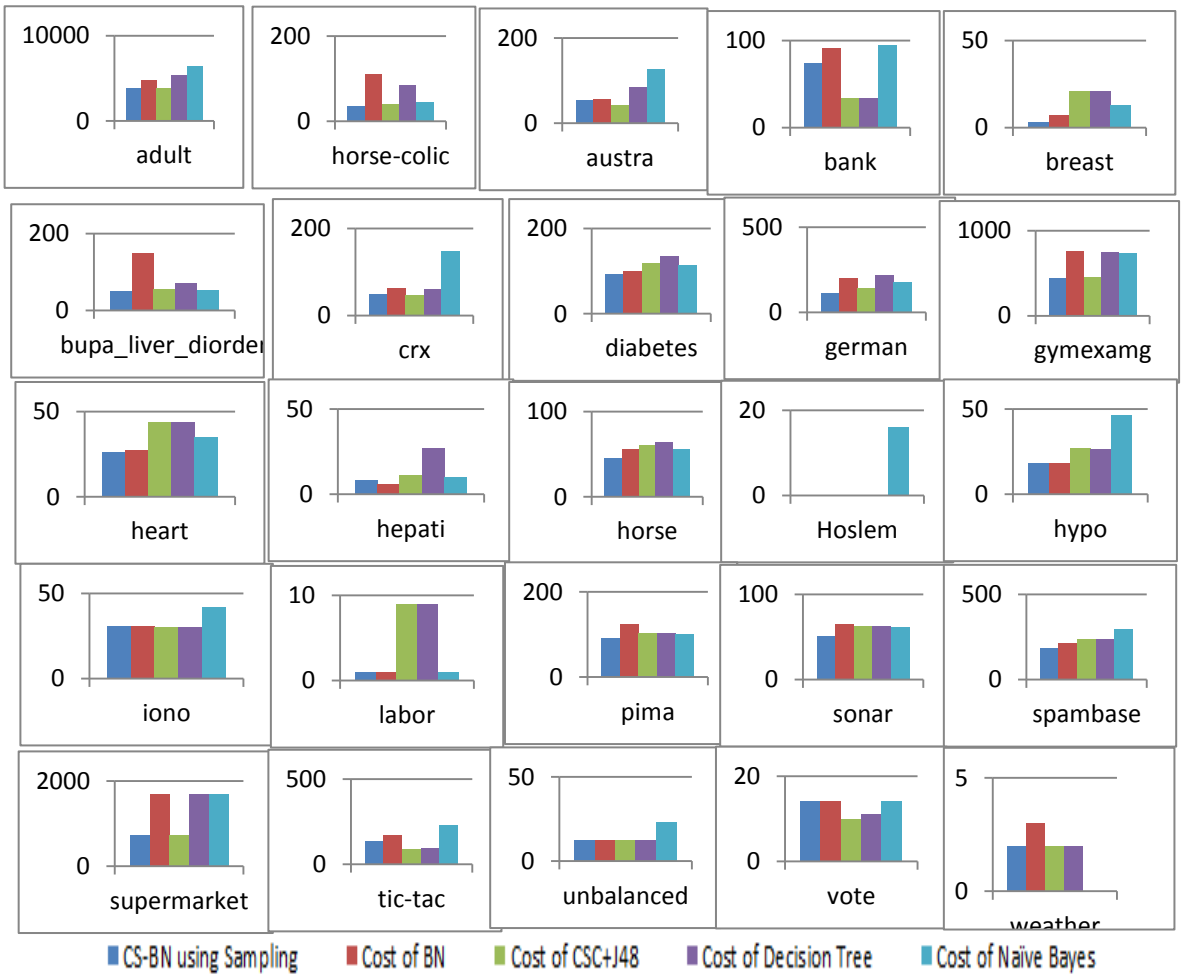

Fig. 4. Expected cost of CS-BN via changing the distributions and existing algorithms

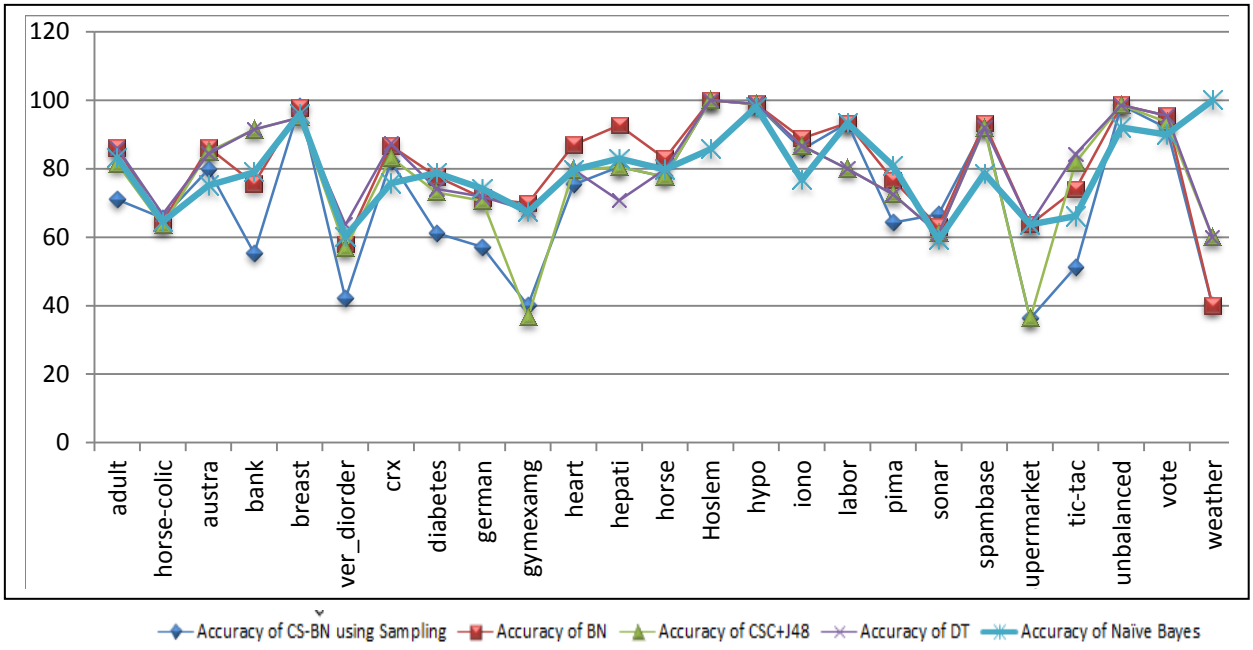

Fig. 5. Accuracy of CS-BN via changing the distributions and existing algorithms.

5 Results and discussion

This experiment shows that the number of misclassifications of rare class (more expensive) are always less than the number of misclassifications for the rare class in the original TAN algorithm for most of the data. Thus, the results are always better in terms of cost, as we can see in Table 1. Furthermore, as shown in Figure 4, for most of the data sets, the changing proportion method (CS-BN via sampling) gives good results compared to the original TAN, MetaCost, Decision Tree (j48), and Naïve Bayes(NB). On the other hand, in Figure 5, it is shown that the accuracy, in most cases, is a little lower than the original TAN algorithm.

As consequence, changing the data distributions before applying TAN classifier yields good results in most data; especially if the data are not very highly skewed to one class. Therefore, the expected cost of using our experiments will provide a reduction of misclassification costs, compared to the original algorithm, which does not use this method. Therefore, we believe that the proposed CS-BN approach of changing the data distributions will produce good results in terms of cost and accuracy.

6 Conclusion

Although much work has been conducted on the development of cost-sensitive decision tree learning, little has been conducted on assessing whether other classifiers, such as Bayesian networks, can lead to better results. Therefore, taking into account work with the folk theorem [22,5], a new Black Box method, based on amending the distribution of examples to reflect the costs of misclassification, was applied in order to develop cost-sensitive Bayesian networks. A preliminary experiment, amending the distributions of TAN, has been carried out on several datasets previously studied by various researchers using different methods.

Our *CS-BN with sampling* approach has been evaluated and compared with MetaCost+J4.8, standard decision tree (J48), standard Bayesian networks approaches, and standard Naïve Bayes(NB . The results for over 25 data sets show that the use of sampling yields better results than the current leading approach; namely, the use of MetaCost+J4.8.

In conclusion, our new CS-BN algorithm has been developed and explored by using a Black Box approach with sampling that amends the data distribution to take account of costs shows promising results in comparison to existing cost-sensitive tree induction algorithms.

REFERENCES

[1] Friedman, Jerome H. "Data Mining and Statistics: What's the connection?." *Computing Science and Statistics* 29, no. 1 (1998):pp. 3-9.

[2] Pearl, Judea. "Embracing Causality in Formal Reasoning." In *AAAI*, pp. 369-373. 1987.

[3] S. Vadera, and D. Ventura, "A Comparison of Cost-Sensitive Decision Tree Learning Algorithms", *Second European Conference in Intelligent Management Systems in Operations*, 3 4 July, University of Salford, Operational Research Society, Birmingham, UK, 2001, pp. 79-

86.
[4] Lomax, Susan, and Sunil Vadera. "A survey of cost-sensitive decision tree induction algorithms." *ACM Computing Surveys (CSUR)* 45, no. 2 (2013): pp 16:1-16:35.
[5] Zadrozny, Bianca, John Langford, and Naoki Abe. "Cost-sensitive learning by cost-proportionate example weighting." In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pp. 435-442. IEEE, 2003a.
[6] Zadrozny, Bianca, John Langford, and Naoki Abe. "A simple method for cost-sensitive learning." IBM Technical Report RC22666, (2003b).
[7] Sheng, Victor S., and Charles X. Ling. "Roulette sampling for cost-sensitive learning." In *Machine Learning: ECML 2007*, pp. 724-731. Springer Berlin Heidelberg, 2007.
[8] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *pp.* 1106.1813 (2011).
[9] Ma, Guang-Zhi, Enmin Song, Chih-Cheng Hung, Li Su, and Dong-Shan Huang. "Multiple costs based decision making with back-propagation neural networks." *Decision Support Systems* 52, no. 3 (2012):pp. 657-663.
[10] Maloof, Marcus A. "Learning when data sets are imbalanced and when costs are unequal and unknown." In *ICML-2003 workshop on learning from imbalanced data sets II*, vol. 2, pp. 2-1. 2003.
[11] Drummond, Chris, and Robert C. Holte. "C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling." In *Workshop on Learning from Imbalanced Datasets II*, vol. pp.11. 2003.
[12] Kubat, Miroslav, and Stan Matwin. "Addressing the curse of imbalanced training sets: one-sided selection." In *ICML*, *In Proceedings of the Fourteenth International Conference on Machine Learning*, vol. 97, pp. 179-186. 1997.
[13] Ling, Charles X., and Chenghui Li. "Data Mining for Direct Marketing: Problems and Solutions." In *KDD*, vol. 98, pp. 73-79. 1998.
[14] Domingos, Pedro. "Metacost: A general method for making classifiers cost-sensitive." In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 155-164. ACM, 1999.
[15] Vadera, Sunil. "CSNL: A cost-sensitive non-linear decision tree algorithm." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 4, no. 2 (2010): 6.
[16] Pazzani, Michael J., Christopher J. Merz, Patrick M. Murphy, Kamal Ali, Timothy Hume, and Clifford Brunk. "Reducing Misclassification Costs." In *ICML*, vol. 94, pp. 217-225. 1994.
[17] Batista, Gustavo EAPA, Ronaldo C. Prati, and Maria Carolina Monard. "A study of the behavior of several methods for balancing machine learning training data." *ACM Sigkdd Explorations Newsletter* 6, no. 1 (2004):pp. 20-29.
[18] Agarwal, Alekh. "Selective sampling algorithms for cost-sensitive multiclass prediction." In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1220-1228. 2013.
[19] Fayyad, Usama, and Keki Irani. "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning." , *Proceedings of the International Joint Conference on Uncertainty in AI* ,(1993): pp. 1022-1027.
[20] Asuncion, Arthur, and David Newman. "UCI machine learning repository." (2007). http://archive.ics.uci.edu/ml/.
[21] Elkan, Charles. "The foundations of cost-sensitive learning." In *International joint conference on artificial intelligence*, vol. 17, no. 1, pp. 973-978. LAWRENCE ERLBAUM ASSOCIATES LTD, 2001.