ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1638

Learning a Bayesian Network Classifier by Jointly Maximizing Accuracy and Information

Dan Halbersberg and **Boaz Lerner** [1]

Abstract. Although recent studies have shown that a Bayesian network classifier (BNC) that maximizes the classification accuracy (i.e., minimizes the 0/1 loss function) is a powerful tool in knowledge representation and classification, this classifier focuses on the majority class, is usually uninformative about the distribution of misclassifications, and is insensitive to error severity (making no distinction between misclassification types). We propose to learn a BNC using an information measure (IM) that jointly maximizes classification and information, and evaluate this measure using various databases. We show that an IM-based BNC is superior to BNCs learned using other measures, especially for ordinal classification and imbalanced problems, and does not fall behind state-of-the-art algorithms with respect to accuracy and amount of information provided.

1 Introduction and Background

Prediction and identification of key factors in imbalance ordinal problems are difficult for several reasons. First, classifiers that maximize accuracy (ACC) during learning do not account for error distribution and, thus, are not informative enough about the classification result. On the other hand, classifiers that account for error distribution usually are not accurate enough. Second, for imbalanced data, classifiers usually predict all minority samples as the majority class. Tackling imbalance by down-sampling the majority class, up-sampling the minority class, or applying different costs to different misclassifications provide an optimistic ACC estimate, and thus are not recommended [4]. Third, 0/1 loss function classifiers are not optimized to tackle different error severities differently; for instance, they consider misclassification of fatal accidents as severe, similar to misclassification of fatal accidents as minor. Fourth, classifiers (e.g., SVM, NN) usually excel in prediction but not in knowledge representation, which is a main goal of this study.

The Bayesian network classifier (BNC) excels in knowledge representation, which makes it ideal to identify key factors as required, but like other classifiers, it suffers from the first three problems. It has been claimed [2] and shown [3] that to achieve high ACC, a BNC should maximize a (discriminative) score which is specific to classification, and not a general inference score based on likelihood. Therefore, to tackle the above concerns, we first consider replacing ACC with four existing scores, each of which accounts for the entire confusion matrix (CM) and not just its diagonal (ACC): 1) *Mutual information* (**MI**) that is defined between two M-dimensional vectors, X and Y, holding predictions and true values for M possible classes, respectively [1]; 2) *Mean absolute error* (**MAE**) that is the average deviation between X and Y; 3) *Matthew correlation coefficient* (**MCC**) that is the correlation between the true (Y) and predicted (X) class matrices [1]; and 4) *Confusion entropy* (**CEN**) [5] that exploits the distribution of misclassifications of a class as any of the $M-1$ other classes.

Second, since none of the above measures accounts for all concerns, we propose a novel information measure (IM) that uses MI to evaluate the error distribution and a factor we introduce to measure error severity (ES) between predictions and true values,

$$IM_\alpha = \sum_x^X \sum_y^Y P(x,y)\left(-log\left(\frac{\alpha P(x,y)}{P(x)P(y)}\right) + log(1+\alpha|x-y|)\right). \tag{1}$$

When $\alpha = 1$, $IM_\alpha = IM$, and ES measures a "classification distance" $|x-y|$, which is transformed using the weighted by the joint distribution logarithm to MI "units". Both $P(X,Y)$ and $|x-y|$ are measured using the CM. When predictions are uniformly distributed (maximum entropy), MI contributes the most to IM. When there is no error between X and Y (off-diagonal elements are 0), the only contribution to IM is from MI, and if, in addition, the classes are balanced, IM is minimized to $-log(M)$. When the error severity is maximal, ES contributes $log(1+M-1) = log(M)$; hence, MI and ES contradict each other, and IM is balanced in $[-log(M), log(M)]$. We seek a classifier whose prediction and true value distributions correspond to each other, while its errors are the least severe. When $\alpha > 1$, α is a user or data-defined constant that balances ACC, information, and ES, IM$_\alpha$ is a generalization of IM. Then, it is easy to show that $IM_\alpha = IM - log(\alpha) \times ACC$, i.e., IM$_\alpha$ is monotonic with ACC, and when $\alpha \to \infty$, $log(\alpha)$ dominates IM$_\alpha$.

To demonstrate the value of these two novel measures in comparison to the existing measures, we conducted experiments with synthetic CMs and summarize the most important properties of the measures regarding whether they: 1) balance ACC and information, 2) prefer balanced class distribution, 3) are sensitive to the error distribution, 4) tackle error severity, and 5) are sensitive to the number of classes. ACC and MCC do not meet any of the above properties; MI meets 1, 2, and 5; CEN meets 1 and 3; and MAE meets only 4. IM and IM$_\alpha$ are the only measures to meet them all.

2 Evaluation, Experiments, and Results

We compared the ability of each of the two proposed measures to augment learning of a state-of-the-art BNC called RMCV [3] with those of the existing measures (ACC, MI, MAE, MCC, and CEN), suggesting seven algorithms (classifiers) for evaluation. In each learning step of the RMCV algorithm, neighboring graphs (edge addition/deletion/reversal) are compared with the current graph as part of a greedy hill climbing, and learning proceeds if the measure computed on the validation set is improved by any of the neighboring

[1] Ben-Gurion University, Israel, emails: halbersb@bgu.ac.il; boaz@bgu.ac.il

Table 1. Mean|std ACC and normalized IM$_\alpha$ values of BNCs learned using seven measures for 23 artificial (ART) and 16 real-world (RL) DBs.

	ACC performance							IM$_\alpha$ performance						
	IM	IM$_\alpha$	MI	CEN	MCC	MAE	ACC	IM	IM$_\alpha$	MI	CEN	MCC	MAE	ACC
ART	75.4|9	**76.2|9**	74.0|10	*46.8|19*	75.0|10	74.4|11	75.4|9	68.5|9	**69.0|9**	67.6|8	*43.7|11*	68.1|9	67.5|10	68.1|9
RL	78.8|15	**79.0|15**	78.6|15	*68.6|22*	78.7|16	78.7|15	78.6|16	66.6|15	**66.9|15**	66.5|15	*54.9|21*	66.6|15	66.5|15	66.2|15

graphs, which then becomes the new current graph. When learning is completed, a CM is computed using the test set. This CM was evaluated using the seven measures. IM$_\alpha$ was normalized using min-max values in order to select best α (we heuristically examined values of α in $[2 : M, M^2, M^3]$ for M classes, and selected the α that maximized IM$_\alpha$ on a validation set independent of the training and test sets). We made the evaluation using artificial (ART) and real-world (RL) DBs (all classification problems were ordinal). In the results reported (Table 1 and Table 2), **bold** and *underlined italic* fonts indicate the best and worst algorithms, respectively.

First, we evaluated the algorithms using 23 ART DBs that were generated from a synthetic 20-node BN structure, in which the class variable Markov blanket includes 4 parents, 3 children, and 3 parents to common children. To test various scenarios and to simulate a broad range of problems, we changed the number of values of the class variable between 2 and 9, and sampled 2,000 samples in DBs 1–8; kept four values to the class variable and changed the number of samples between 500 and 3,000 in DBs 9–14; and sampled 2,000 samples and kept four classes, but changed their prior probabilities to represent different degrees of imbalance – from pure balance, through different levels of imbalance, to very high imbalance – in DBs 15–23. Each DB was re-sampled to create ten data permutations, and each permutation was divided into five folds (i.e., CV5). Table 1 (top) shows evaluation of the seven algorithms averaged over the 23 ART DBs according to the ACC (left) and IM$_\alpha$ (right). BNCs learned based on the IM$_\alpha$ measure (1) perform better than BNCs learned based on all other measures, regardless if the evaluation is based on the IM$_\alpha$ measure or ACC, which is interesting to see because classifiers trained to maximize the IM$_\alpha$ measure are not expected to also maximize ACC. BNCs based on ACC, MCC, or IM, are behind, and those learned based on the CEN measure provide the poorest performance. The reason that the CEN-based BNC performs so poorly is because empty graph initialization of the RMCV algorithm creates an initial CM that has entries only for one class, the majority class, leaving the total CEN measure relatively low (better). Thus, it stops at this local minimum and cannot proceed further.

To check if these differences are statistically significant, we performed a Friedman non-parametric test followed by a Nemenyi post-hoc test. Table 2 (top) shows the average ranks of the algorithms (lowest is best) according to ACC and IM$_\alpha$, based on the Friedman test. The Nemenyi test shows (with a 0.05 confidence level) that all algorithms are significantly better than CEN with respect to the ACC and IM$_\alpha$ measures, and the IM$_\alpha$-based BNC is significantly superior to those based on MI, MCC, MAE, and ACC. In addition, the BNC-IM$_\alpha$ has significantly better average ranks than the other algorithms have regardless of the measure that evaluates performance.

Table 2. Average ranks according to ACC and IM$_\alpha$ of BNCs learned using seven measures for the ART and RL DBs.

		IM	IM$_\alpha$	MI	CEN	MCC	MAE	ACC
AR	ACC	3.2	**1.6**	4.7	*6.9*	3.9	3.8	3.8
	IM$_\alpha$	3.0	**1.7**	4.3	*6.9*	4.0	4.0	4.1
RL	ACC	2.6	**2.1**	4.1	*6.0*	4.1	4.3	4.8
	IM$_\alpha$	2.6	**2.1**	3.8	*6.1*	4.4	4.1	4.9

Next, we extended the evaluation of the measures using 14 UCI RL DBs (Australian, Autombp, Bostonhousing, Car, Cleve, Corral, Glass, Hepatitis, Machinecpu, Mofn, Mushroom, Shuttle, Stocksdomain, and Voting) and two of our own DBs: Amyotrophic lateral sclerosis (ALS) and Missed due date. The problems represented by these 16 DBs have 2–10 classes, 7–29 variables, 80–10,500 samples, and different degrees of class imbalance, posing a range of challenges to the classifiers. Again, ten random permutations were made to each DB, which were used over a CV5 experiment. Table 1 (bottom) shows the evaluation of the seven algorithms according to ACC (left) and IM$_\alpha$ (right) performances. Once again, IM$_\alpha$-based BNCs preform better on average than BNCs learned based on all other measures. Table 2 (bottom) shows the average ranks according to the Friedman test according to ACC and IM$_\alpha$. The results are consistent with those of the ART DBs, showing that IM$_\alpha$ is ranked first, followed by IM, MCC, and MAE. The ACC-based BNC was the second worst classifier, which re-emphasizes the motivation to replace it. According to Nemenyi post-hoc test (with 0.05 confidence level), BNC-IM$_\alpha$ is superior to BNCs-MI, CEN, MCC, and ACC with respect to ACC and to BNCs-CEN, MCC, and ACC with respect to IM$_\alpha$. In addition, we expended our evaluation to other state-of-the-art algorithms suitable for ordinal classification, such as ordinal regression and ordinal DT with a cost matrix equivalent to that used by IM and IM$_\alpha$. Friedman and Nemenyi tests found BNC-IM$_\alpha$ to be superior to these algorithms.

3 Summary and Discussion

Learning by only maximizing ACC and ignoring the error distribution and severity in class-imbalance problems results in accurate classification of only the major class at the expense of incorrect prediction of the minor one. We proposed an information measure, IM, and a weighted version of it, IM$_\alpha$, to tackle these limitations in ordinal classification problems. We implemented them as a discriminative score in an algorithm for learning a BNC and demonstrated their advantage compared to other measures. IM and IM$_\alpha$ are specifically suited to any imbalance ordinal classification problem. If a problem is not ordinal, the contribution and impact of the ES term in the measures will vanish, and with no imbalance, the measures advantage over others may decrease. For problems with high imbalance and error that account differently for different classes, the advantage of these measures over other ACC and information measures is large.

REFERENCES

[1] P. Baldi, et al., 'Assessing the accuracy of prediction algorithms for classification: an overview', *Bioinformatics*, **16**(5), 412–424, (2000).

[2] N. Friedman, et al., 'Bayesian network classifiers', *Machine Learning*, **29**(2-3), 131–163, (1997).

[3] R. Kelner and B. Lerner, 'Learning Bayesian network classifiers by risk minimization', *International Journal of Approximate Reasoning*, **53**(2), 248–272, (2012).

[4] F. Provost, 'Machine learning from imbalanced data sets', in *Proceedings of the AAAI Workshop on Imbalanced Data Sets*, pp. 1–3, (2000).

[5] J. M. Wei, et al., 'A novel measure for evaluating classifiers', *Expert Systems with Applications*, **37**(5), 3799–3809, (2010).