A Comparison of Two Approaches to Data Mining from Imbalanced Data

Jerzy W. Grzymala-Busse[1], Jerzy Stefanowski[2], and Szymon Wilk[2]

[1] Department of Electrical Engineering and Computer Science,
University of Kansas, Lawrence, KS 66045, USA and
Institute of Computer Science, Polish Academy of Sciences, 01-237 Warsaw, Poland
Jerzy@ku.edu http://lightning.eecs.ku.edu/index.html
[2] Institute of Computing Science, Poznan University of Technology,
60-965 Poznan, Poland
{Jerzy.Stefanowski, Szymon.Wilk}@cs.put.poznan.pl

Abstract. Our objective is a comparison of two data mining approaches to dealing with imbalanced data sets. The first approach is based on saving the original rule set, induced by the LEM2 algorithm, and changing the rule strength for all rules for the smaller class (concept) during classification. In the second approach, rule induction was split: the rule set for the larger class was induced by LEM2, while the rule set for the smaller class was induced by EXPLORE, another data mining algorithm. Results of our experiments show that both approaches increase the sensitivity compared to the original LEM2. However, the difference in performance of both approaches is statistically insignificant. Thus the appropriate approach to dealing with imbalanced data sets should be selected individually for a specific data set.

1 Introduction

During data mining from real-life data, sizes of classes (concepts) are frequently different. Quite often the class which is critical from the domain point of view (the primary class) includes a much smaller number of cases while other (secondary) classes form the majority of cases [6]. This situation is typical in medical problems, where the task is to diagnose a specific disease. The primary class usually describes patients requiring special attention while all remaining cases are members of the secondary class (e.g., healthy patients). Similar situations also occur in other domains, e.g., in financial analysis of loan policy or bankruptcy.

Standard classifiers derived from such data sets are also affected by a lack of balance. That is, their predictive accuracy is biased towards majority classes and they usually have difficulties with correct classification of cases from the primary classes. Since the primary class is more important, costs of false positives and false negatives may drastically differ. Using again an example of medical diagnosis, the total classification accuracy is misleading as an indicator of the classifier quality for imbalanced data. Diagnosis is characterized by sensitivity (the conditional probability of the set of correctly classified cases from the primary class,

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3213, pp. 757–763, 2004.

given the primary class) and by specificity (the conditional probability of the set of correctly recognized cases from the secondary class, given the secondary class). In such applications more attention is given to sensitivity than to specificity.

In our research we tested two approaches to increasing the sensitivity of the primary class for rule-based classifiers. In both approaches, initial rules were induced by the LEM2 algorithm. An original version of LEM2 induces a minimal set of rules from rough approximations of classes [2], [3]. Generated rules are then used by the LERS (Learning from Examples based on Rough Sets) "bucket brigade" classification strategy. The first technique to improve sensitivity is based on increasing strengths of rules describing the primary class. The rule strength is defined as the number of training cases correctly classified by the rule. The idea is to multiply the strengths of all primary class rules by the same real number, called strength multiplier, while not changing the strength of rules from the secondary classes. As a result, during classification of new cases, such primary class rules have an increased chance to classify these cases as being members of the primary class.

The second technique is based on a different principle. A minimal set of rules for the primary class is replaced by a new set of rules with the strength greater than a certain threshold. Such rules are discovered by a special algorithm, called EXPLORE [8]. If the strength threshold is sufficiently low, EXPLORE may generate much more rules then LEM2. Thus, by using such rules for the primary class, while preserving the original set of rules for the secondary class, the chance that a case from the primary class is selected by a classifier is increased and sensitivity should improve.

The main aim of this study is to evaluate the performance of both techniques on several imbalanced data sets. Moreover, we compare both techniques using a standard scheme of applying LEM2 with LERS classification strategy.

2 Data Mining with LERS

Both presented approaches to some extent employ the LEM2 algorithm which uses rough set theory for inconsistent data. LEM2 is a component of the LERS data mining system [2], [3]. In rough set theory inconsistencies are not removed from consideration. Instead, lower and upper approximations of the concept are computed. On the basis of these approximations, two corresponding sets of rules: certain and possible, are induced.

In our experiments we used the LERS version of the classification system. For classification of unseen cases system LERS employees a modified "bucket brigade algorithm". In this approach, the decision to which concept a case belongs is made using two factors: *strength* and *support*. In LERS, the strength is the total number of cases correctly classified by the rule during training. The second factor, *support*, is related to a concept and is defined as the sum of strengths of all matching rules from the concept. The concept receiving the largest support wins the contest. This process remains voting by rules for concepts.

3 Sensitivity and Specificity

In many applications, e.g., in medicine, we distinguish between two classes: primary and secondary. The primary class, more important, is defined as the class of all cases that should be diagnosed as affected by a disease. The set of all correctly classified cases from the primary class are called true-positives, incorrectly classified primary cases are called false-negatives, correctly classified secondary cases are called true-negatives, and incorrectly classified secondary cases are called false-positives.

Sensitivity is the conditional probability of true-positives given primary class, i.e., the ratio of the number of true-positives to the sum of the number of true-positives and false-negatives. Specificity is the conditional probability of true-negatives given secondary class, i.e., the ratio of the number of true-negatives to the sum of the number of true-negatives and false-positives.

Usually, by applying techniques described later, we may increase sensitivity at the cost of specificity. It is difficult to estimate what are the optimal values of sensitivity and specificity. In our experiments we applied an analysis presented in [1]. Let p be a probability of the correct prediction, i.e., the ratio of all true positives and all false positives to the total number of all cases. Let P be the probability of an actual primary class, i.e., the ratio of all true positives and all false negatives to the total number of all cases. Then

$$p = Sensitivity * P + (1 - Specificity) * (1 - P).$$

Following [1], we would like to see the change in p as large as possible with a change in P, i.e., we would like to maximize

$$\frac{dp}{dP} = Sensitivity + Specificity - 1.$$

Thus the optimal values of sensitivity and specificity correspond to the maximal value of $Sensitivity + Specificity$. The sum of sensitivity and specificity is called a *gain*. Thus, in our experiments the objective was to maximize gain.

4 Increasing the Strength of Rules

As a result of rule induction, the average of all rule strengths for the bigger class is greater than the average of all rule strengths for the more important but smaller primary class. During classification of unseen cases, rules matching a case and voting for the primary class are outvoted by rules voting for the bigger, secondary class. Thus the sensitivity is low and the resulting classification system would be rejected by the users.

Therefore it is necessary to increase sensitivity. The simplest way to increase sensitivity is to add cases to the primary class in the data set, e.g., by adding duplicates of the available cases. The total number of training cases will increase, hence the total running time of the rule induction system will also increase.

Adding duplicates will not change the knowledge hidden in the original data set, but it may create a balanced data set so that the average rule set strength for both classes will be approximately equal. The same effect may be accomplished by increasing the average rule strength for the primary class. In our first approach to dealing with imbalanced data sets we selected the optimal rule set by multiplying the rule strength for all rules describing the primary class by the same real number called a *strength multiplier* [4], [5].

In general, the sensitivity increases with the increase of the strength multiplier. At the same time, the specificity decreases. In our experiments, rule strength for all rules describing the primary class was increased incrementally. The process was terminated when gain was decreased.

5 Replacing Rules

Unlike the previous technique, this approach is based on replacing the rule set for the primary class by another rule set, generated directly from data, that improves the chance of the "bucket brigade" algorithm selecting a case from the primary class, as a new case can be matched by multiple rules voting for the primary class.

In order to generate additional rules for the primary class, we apply the EXPLORE algorithm [8]. As opposed to LEM2, EXPLORE induces all rules that satisfy certain requirements, e.g., the strength greater than a given value, or the length of a rule smaller than a specified threshold. The main part of the algorithm is based on the breadth-first search, where rules are generated from the shortest to the longest. Creation of a rule stops as soon as a rule satisfies the requirements or it is impossible to fulfill the requirements in further steps.

Although as mentioned above, there are several requirements that can be specified for EXPLORE, we are focused only on the minimal strength of a rule (for discussion see [7], [8]). The threshold is modified in order to obtain an optimal set of rules, i.e., leading to the best classification outcome [9]. To avoid repeating induction of rules with varying strengths, a set of rules is generated only once for the smallest acceptable threshold, and then appropriate subsets are selected. The smallest strength is set to the minimal strength observed for rules generated for the primary class by LEM2. Rules for the secondary class are created as previously, using LEM2.

To find an optimal set of rules according to the gain criterion described in Section 3 we verify, in a number of steps, various subsets of rules for the primary class, starting from the strongest rules to all rules created by EXPLORE. In each step we consider rules for the primary class with strength greater than the current threshold and combine them with rules obtained for the secondary class into a final set used by the classifier. If the number of rules for the primary class exceeds a number of rules for the class generated by LEM2, we finish the process of finding optimal rules. When the process is completed, we select the threshold and a set of rules leading to the best classification outcome.

6 Experiments

Some of the original data sets, used for our experiments, contained numerical attributes. These attributes were discretized using cluster analysis. Clusters were first formed from data with numerical attributes. Then those clusters were projected on the attributes that originally were numerical. The resulting intervals were merged to reduce the number of intervals and, at the same time, to preserve consistency.

Some data sets contained missing attribute values, which were substituted with the most frequent value among cases belonging to the considered class.

For calculation of classification performance we used two fold cross validation. For both approaches we used the same sets of cases, with the same split into two subsets. Though two-fold cross validation may be not sufficient to estimate the actual error rate, our objective was to compare our approaches to handling imbalanced data sets.

Most of the data sets, presented in Table 1, were taken from the Repository at the University of California, Irvine, CA. Others come from medical applications of rule induction approaches [10]. In Tables 2–4, sensitivity, specificity, gain and the total error are presented.

Table 1. Data sets used in experiments

	Number of cases			Ratio of cases	
Data set	Total	Primary	Secondary	Primary	Secondary
ABDOMINAL-PAIN	723	202	521	27.9%	72.1%
BREAST-SLOVENIA	294	89	205	30.3%	69.7%
BREAST-WISCONSIN	625	112	513	17.9%	82.1%
BUPA	345	145	200	42.0%	58.0%
GERMAN	666	209	457	31.4%	68.6%
HEPATITIS	155	32	123	20.6%	79.4%
PIMA	768	268	500	34.9%	65.1%
SCROTAL-PAIN	201	59	142	29.4%	70.6%
UROLOGY	498	155	343	31.1%	68.9%

7 Conclusions

Results of our experiments show that an increase in gain, comparing with the original LEM2, may be accomplished by both approaches: changing strength multipliers for rules describing the primary class and by replacing rule sets for the primary class using EXPLORE.

The purpose of our experiments was to compare both approaches to dealing with imbalanced data sets. In order to compare the overall performance of both approaches, the Wilcoxon Signed Ranks Test, a nonparametric test for significant differences between paired observations, was used. As a result, the difference in performance for both approaches to dealing with imbalanced data sets, in terms

Table 2. Results for the original LEM2 algorithm

Data set	Sensitivity	Specificity	Gain	Error
ABDOMINAL-PAIN	0.5842	0.9290	1.5132	16.74%
BREAST-SLOVENIA	0.3647	0.8856	1.2503	26.92%
BREAST-WISCONSIN	0.3125	0.9259	1.2384	18.40%
BUPA	0.3241	0.7400	1.0641	43.48%
GERMAN	0.3014	0.8468	1.1482	32.43%
HEPATITIS	0.4375	0.9512	1.3887	15.48%
PIMA	0.3918	0.8260	1.2178	32.55%
SCROTAL-PAIN	0.5424	0.8310	1.3734	25.37%
UROLOGY	0.1218	0.8227	0.9445	39.60%

Table 3. Best results for increasing rule strength

Data set	Multiplier	Sensitivity	Specificity	Gain	Error
ABDOMINAL-PAIN	5.0	0.8069	0.8484	1.6553	16.32%
BREAST-SLOVENIA	1.0	0.3647	0.8856	1.2503	26.92%
BREAST-WISCONSIN	5.0	0.5714	0.8674	1.4388	18.56%
BUPA	3.0	0.5586	0.5850	1.1436	42.61%
GERMAN	4.0	0.5789	0.6411	1.2200	37.84%
HEPATITIS	18.0	0.8438	0.7724	1.6162	21.29%
PIMA	3.5	0.5933	0.7640	1.3573	29.56%
SCROTAL-PAIN	3.0	0.6780	0.8099	1.4879	22.89%
UROLOGY	14.0	0.5192	0.4942	1.0134	49.48%

Table 4. Best results for replacing rules (EXPLORE approach)

Data set	Support	Sensitivity	Specificity	Gain	Error
ABDOMINAL-PAIN	16.0	0.6939	0.9175	1.6114	14.52%
BREAST-SLOVENIA	3.0	0.4709	0.8411	1.3120	26.92%
BREAST-WISCONSIN	2.0	0.6385	0.8160	1.4545	21.43%
BUPA	2.0	0.4275	0.6300	1.0575	45.50%
GERMAN	5.0	0.6271	0.7265	1.3536	30.50%
HEPATITIS	6.0	0.5830	0.9175	1.5005	15.52%
PIMA	3.0	0.5686	0.7829	1.3514	29.30%
SCROTAL-PAIN	4.0	0.6887	0.8724	1.5611	18.44%
UROLOGY	6.0	0.3403	0.7017	1.0420	41.57%

of gain, is statistically insignificant. Additionally, the same conclusion is true for the error rate: the difference in performance for both approaches, in terms of error rate, is also statistically insignificant. Therefore, the appropriate approach to dealing with imbalanced data sets should be selected individually for a specific data set. The first approach to increasing sensitivity, based on changing the rule strength for the primary class, is less expensive computationally than the second approach, based on replacing the rule set for the primary class.

We can extend both approaches by also post-processing rule sets for stronger secondary class using rule truncation, i.e., removing weak rules describing only a few training cases. Such possibilities can be explored in further research.

For many important applications, e.g., medical area, an increase in sensitivity is crucial, even if it is achieved at the cost of specificity. Thus, the suggested approaches to dealing with imbalanced data sets may be successfully applied for data mining from imbalanced data.

Acknowledgment. This research was partially supported by the State Committee for Research (KBN) of Poland, grant 3 T11C 050 26.

References

1. Bairagi, R., and Suchindran, C. M.: An estimator of the cutoff point maximizing sum of sensitivity and specificity. *Sankhya, Series B, Indian Journal of Statistics* **51** (1989) 263–269.
2. Grzymala-Busse, J.W.: LERS—a system for learning from examples based on rough sets. In: Slowinski, R. (ed.): Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory. Kluwer Academic Publishers, Dordrecht, Boston, London (1992) 3–18.
3. Grzymala-Busse, J.W.: A new version of the rule induction system LERS. *Fundamenta Informaticae* **31** (1997) 27–39.
4. Grzymala-Busse, J.W., Goodwin, L.K., Grzymala-Busse, W.J., Zheng, X.: An approach to imbalanced data sets based on changing rule strength. Learning from Imbalanced Data Sets, AAAI Workshop at the 17th Conference on AI, AAAI-2000, Austin, TX, July 30–31, 2000, 69–74.
5. Grzymala-Busse, J. W., Goodwin, L. K., and Zhang, X.: Increasing sensitivity of preterm birth by changing rule strengths. Proceedings of the Eigth Workshop on Intelligent Information Systems (IIS'99), Ustron, Poland, June 14–18, 1999, 127–136.
6. Japkowicz, N.: Learning from imbalanced data sets: a comparison of various strategies. Learning from Imbalanced Data Sets, AAAI Workshop at the 17th Conference on AI, AAAI-2000, Austin, TX, July 30–31, 2000, 10–17.
7. Stefanowski J.: On rough set based approaches to induction of decision rules. In: Skowron, A. and Polkowski L. (eds): Rough Sets in Knowledge Discovery, Physica Verlag, Heidelberg (1998) 500–529.
8. Stefanowski J., Vanderpooten D.: Induction of decision rules in classification and discovery-oriented perspectives. *International Journal of Intelligent Systems* **16** (2001), 13–28.
9. Stefanowski J., Wilk S.: Evaluating business credit risk by means of approach integrating decision rules and case based learning. *International Journal of Intelligent Systems in Accounting, Finance and Management* **10** (2001) 97–114.
10. Wilk S., Slowinski R., Michalowski W., Greco S.: Supporting triage of children with abdominal pain in the emergency room. *European Journal of Operation Research* (in press).