Jerzy STEFANOWSKI
POLITECHNIKA POZNAŃSKA, INSTYTUT INFORMATYKI

Selective filtering for learning classifiers from imbalanced data

Dr hab. inż. Jerzy STEFANOWSKI

Zainteresowania naukowe obejmują problematykę systemów uczących się (ang. machine learning), odkrywania wiedzy w bazach danych i eksploracji danych (ang. data mining), inteligentnych systemów wspomagania decyzji oraz modelowania niespójności danych z wykorzystaniem teorii zbiorów przybliżonych. Ukończył Wydział Elektryczny Politechniki Poznańskiej, gdzie obronił doktorat (1994) oraz uzyskał stopień doktora habilitowanego (2001).

e-mail: Jerzy.Stefanowski@cs.put.poznan.pl

Abstract

This paper concerns problems of automatic learning rule based classifiers from imbalanced data, where the minority class of primary importance is underrepresented in comparison to majority classes. To improve recognition of the minority class, we present the new approach, where the rule induction is combined with the selective filtering phase that removes noisy and borderline majority class examples from the input data. This approach is evaluated in a comparative experimental study.

Keywords: machine learning, classifiers, imbalanced data.

Selektywny wybór przykładów w konstrukcji klasyfikatorów z niezrównoważonych danych

Streszczenie

W artykule omawia się problemy automatycznego konstruowania klasyfikatorów, będących zbiorem reguł decyzyjnych, z niezrównoważonych danych, w których klasa obiektów, będących przedmiotem szczególnego zainteresowania, zawiera zdecydowanie mniej przykładów niż inne klasy. W celu polepszenia zdolności rozpoznawania przykładów z klasy mniejszościowej przedstawia się propozycje wykorzystania selektywnego wyboru przykładów z klasy większościowej przed fazą indukcji reguł. Podejście jest ocenione w eksperymentach porównawczych ze innymi metodami.

Słowa kluczowe: systemy uczące się, klasyfikacja, niezrównoważone dane.

1. Introduction

Machine learning and data mining are areas of growing research interest. One of their common tasks is *supervised learning*, which aims at discovering from historical data a representation of knowledge that assigns examples, each described by a fixed set of attributes, to known a priori classes. Such a *classification knowledge* derived by an algorithm from learning examples can be successively used to classify new objects. In this sense learning process results in creating a *classifier* [9].

There are several aspects that might cause difficulties for a learning algorithm and decrease performance of learned classifiers. One of them is related to *class imbalance* in the input data, i.e. to a situation when one class (further called the *minority class*) includes much smaller number of examples comparing to other classes [1, 3, 11]. Such data could be met in practice as some processes produce certain observations with a different frequency. A good example is medicine, where databases regarding a rare, but dangerous, disease usually contain a smaller group of patients requiring a special attention while there other classes contain much higher number of patients [4]. Similar situations occur, e.g. in technical diagnostics or continuous fault-monitoring tasks, information retrieval [3, 6, 11].

The total *classification accuracy* (an average percentage of all testing examples correctly recognized by the classifier) is not the only and the best criterion characterizing the classifier performance for such data sets. The users prefer high enough recognition of the minority class and the final decision is characterized rather by its *sensitivity* (the ratio of correctly recognized examples from the minority class) and its *specificity* (the ratio of correctly excluded examples from other classes).

The high imbalance between classes is reported to be an important obstacle in inducing classifiers. Their performance is often degraded as they are biased towards recognition of majority class examples and they usually have difficulties to classify correctly new objects from the minority class [11]. Researchers also indicate other aspects of imbalance data, e.g. the minority class may overlap heavily the majority classes, i.e., there is no clear boundary between them [6]. Boundary region may be affected by ambiguous examples from other classes, which may lead to incorrect classification of examples from the minority class.

In recent years the problem of dealing with the class imbalance has received a growing research interest from the machine learning and data mining communities. Although several methods have been proposed, see e.g. their review in [1, 11], the research problem is still open.

The author with co-operators also introduced an approach which *modifies* the rule *classifier structure* to increase its sensitivity for recognizing the minority class examples [5]. We focused our interest on generating a larger rule set of the minority class, while inducing minimal sets of rules for other, majority classes. As a result of modifying rules we increased the chance of predicting the minority class during the classification strategy for new objects.

The other direction of improving the recognition of the minority class is to focus our attention on a *preprocessing stage* before inducing a classifier, e.g. by appropriate sampling, which could transform the original class distribution into more balanced one – some other researchers also undertaken similar research [6]. However, it should not be just a simple class balancing, e.g. by random duplicating several minority class examples. It may be beneficial to focus attention on *noisy* majority class examples or *boundary* examples between classes as they are crucial for classifying examples from imbalanced classes.

Therefore, we propose to perform a kind of *selective filtering*, where these examples are deleted. We hope that this type of cleaning may give a chance for inducing less specific classification rules and it should also help in a classification phase. In this paper we introduce such an approach, where the selective filtering of these examples is used as a preprocessing stage before inducing rule classifiers by the MODLEM algorithm. The other contribution of this paper is an experimental comparison of this approach against the standard rule based classifiers and two other popular sampling techniques.

2. Related works

We briefly describe only these preprocessing methods, which are the most related to this paper. For reviews of other works, the reader can consult [1, 3, 11].

The sampling in the pre-processing phase transforms the original class distribution into a more balanced and as their result induction of classifiers is less biased to particular classes. The basic approaches include either random *over-sampling* or *under-sampling*. In the former approach the minority class examples are randomly replicated until a balance with cardinalities of majority classes is obtained. Random under-sampling goes in the opposite way - the majority class examples are randomly eliminated until obtaining the same cardinality as the minority class. However, it is

claimed that random under-sampling can discard potentially useful majority class examples that could be valuable for learning a good classifier. On the other hand, simple over-sampling introduces copies of original examples only, which may lead to overspecialization of a classifier. Thus, several more "focused" heuristic techniques have also been introduced.

An example of such focused *under-sampling* is an approach called *one-side-sampling* [6], where the borderline and noisy examples from the majority class are assumed to be a main source of misclassification for minority class examples. Besides an obvious interpretation of noise, borderline examples are treated to be unsafe since a small amount of noise could make them fall on the wrong side of the decision border between classes. These examples are detected by means of, so called, Tomek links [6] and removed. Another approach to removing noisy and borderline examples is *Neighborhood Cleaning Rule* introduced by Laurikkala in [7]. It is based on the Wilson's Edited Nearest Neighbor Rule [12] and removes these majority class examples whose class labels differ from the class of at least two of its three nearest neighbors. Experimental studies [1] showed that both above approaches provide better sensitivity and not worse total accuracy than a simple random over-sampling.

To modify over-sampling, Chawla et al. proposed a heuristic technique, called SMOTE, which over-samples the minority class by creating new synthetic examples [3]. Its main idea is to create these new examples by interpolating several minority class examples that are close one to another. It widens decision boundaries for the minority class. Several experimental results indicate that SMOTE is often more efficient than other sampling methods. Its mixture with elements of under-sampling may also improve the ability to predict the minority class.

The other approaches, e.g. concerning modification of learning or using classifiers, are described in [4, 5, 10, 11].

3. Combining induction of rule classifiers with selective filtering

Following the motivations presented in the previous sections we would like to detect *noisy* or *borderline majority class examples*, which may cause errors while classifying objects from the minority class. They will be removed in a data filtering before inducing rules and finally constructing the classifier.

We implemented a filtering phase inspired by the Laurikkala's work [7]. It cleans majority class examples on the basis of the Wilson's Edited Nearest Neighbor Rule, which recommends removing these examples whose class labels differ from the class of at least two of its three nearest neighbors. This helps to identify noise examples. As it is also necessary to remove borderline examples, the filtering procedure has two stages checking of nearest neighbors, what is summarized below:

- Split learning set E into a minority class C and the rest of data R.
- Identify noisy majority examples from R, i.e. for each example in $e_i \in R$ check: if the classification given by three nearest neighbors of e_i contradicts its original class, then add it to the set A_1.
- For each minority class example $e_j \in C$: if its three nearest neighbors misclassify e_j, then the nearest neighbors that belong to the majority classes are added to the set A_2.
- Remove from the learning set E these majority class examples that belong to a set $A_1 \cup A_2$.

The nearest neighbors of a given example are found as in k-NN algorithm, where k = 3, using a proper distance metric. As the Euclidean distance is not the sufficient for solving real world problems with mixed data described by numeric and nominal attributes, we used the *heterogeneous value difference metric*, which is defined as:

$$HVDM(x,y) = \sqrt{\sum_{a=1}^{m} d_a^2(x_a, y_a)} \quad (1)$$

where $d(x_a,y_a)$ is the distance for attribute a describing examples x, y. For numeric attributes it is defined as normalized absolute value of the distance between values of an attribute. A distance for a nominal attribute is the value difference metric, introduced by Stanfill and Waltz, i.e. for attribute a, its values x_a and y_a it is defined as:

$$vdm_a(x,y) = \sum_{c=1}^{K} (N_{a,x,c} / N_{a,x} - N_{a,y,c} / N_{a,y}) \quad (2)$$

where $N_{a,x}$ is the number of examples where attribute a gets value x_a; $N_{a,x,c}$ is the number of examples where attributes has value x_a and the output class was c. Similar notation refers to value y_a. The distance metric HVDM provides an appropriate normalization between numeric and nominal attributes, as well as between numeric attributes of different scales. Moreover, it handles unknown attribute values by assigning them a large distance.

After removing noisy and borderline examples from the majority classes R in the above filtering procedure, the rule set is induced from remaining data. We decided to use the algorithm MODLEM, which was introduced by Stefanowski [8]. Due to the size of this paper we skip the formal presentation of this algorithm and we only discuss its main idea – for more details see also [9]. It is based on the scheme of a *sequential covering* and it heuristically generates a *minimal set* of decision rules for every decision concept. While looking for the best elementary conditions the entropy based criterion is applied. The extra specificity of the MODLEM algorithm is handling directly numerical attributes during rule induction while elementary conditions of rules are created, without any preliminary *discretization* phase [9].

Finally, the set of induced rules is applied to classify examples using the classification strategy introduced by Grzymala-Busse in LERS system, which takes into account strength of all rules completely matched and also allows partially matches if no rule fits the description of the tested example.

4. Experiments

The usefulness of the proposed approach will be evaluated experimentally. We decided to compare its classification performance against other methods:

- The standard rule based classifier induced by MODLEM algorithm without any additional techniques for handling imbalanced data.
- The simple random under-sampling used together with the rule induction by MODLEM algorithm.
- The simple random over-sampling used together with the rule induction by MODLEM algorithm.

For running our experiments we used our own implementation of MODLEM rule induction algorithm and pre-processing / sampling modules - prepared for the Weka toolkit. Weka is an open source tool containing many machine learning algorithms and data mining methods. This project was started by Witten and Frank; see the WWW link: www.cs.waikato.ac.nz/ml/weka.

In the experiments we calculated two main measures characteristic for studies on imbalanced data, i.e. *sensitivity* and *specificity* – they were calculated for a minority class, being also a class of particular interest in the given problems. Both measures are represented as a number from the interval [0,1], having the following interpretation – the higher value, the better. Furthermore, we will report a total accuracy – as we want to control the overall recognition of other classes besides the minority one. The typical way of representing accuracy is by using percentages – the higher value, the better. The values of all measures are evaluated according to the standard 10-fold stratified cross validation way.

All classifiers were evaluated on 7 data sets, which are popular machine learning benchmarks coming from the UCI repository [2]. The medical data are as follows: women *breast cancer* data coming from Slovenia and other *breast cancer* data coming from *Wisconsin*, *bupa* – live disorders, *pima*, *hepatitis*. Two other data sets are *ecoli* and *glass* recognition. Due to limited paper size we skip detailed characteristic – the reader can find it at [2] These data were chosen to be consistent with other selective sampling studies [1, 6, 7] and on the other hand to consider different degrees of imbalance or to solve difficult classification problems as medical ones. Some of the considered data sets were originally composed of more than two decision classes, however, to simplify problems we decided to group all majority classes into one. Unlike our previous experiments [5,10] we analyzed the original form of data, i.e. they were neither pre-discretized nor missing values were substituted. The obtained results are presented in table 1.

Tab. 1. Classification performance of standard rule classifiers and combined with: simple under-sampling, over-sampling and the new filtering approach

Data set	Classifier type	Minority class sensitivity	Minority class specificity	Total accuracy
Breast cancer Slovenia	standard	0.3056	0.8505	69%
	under-s	0.5971	0.5915	59%
	over-s	0.4043	0.8657	73%
	filtering	0.6264	0.5317	56%
Bupa	standard	0.7290	0.5450	62%
	under-s	0.6707	0.6910	68%
	over-s	0.5935	0.7521	69%
	filtering	0.8767	0.3250	56%
Ecoli	standard	0.4167	0.9667	91%
	under-s	0.8208	0.8430	84%
	over-s	0.5150	0.9578	91%
	filtering	0.7750	0.9335	92%
Glass	standard	0.2500	0.9847	92%
	under-s	0.7800	0.6351	65%
	over-s	0.4050	0.9817	94%
	filtering	0.4000	0.9645	92%
Pima	standard	0.4962	0.8460	72%
	under-s	0.7093	0.7150	71%
	over-s	0.5519	0.8148	72%
	filtering	0.8098	0.6420	70%
Breast cancer Wisconsin	standard	0.9083	0.9586	94%
	under-s	0.9521	0.9484	95%
	over-s	0.8326	0.8619	85%
	filtering	0.9625	0.9652	96%
Hepatitis	standard	0.4833	0.9229	83%
	under-s	0.7372	0.7126	72%
	over-s	0.5447	0.8541	81%
	filtering	0.6500	0.8364	80%

5. Discussion

Let us summarize the results of our experiments. Due to specific properties of imbalance problem, we are the most interested in obtaining the highest values of mainly sensitivity measure. It is also desired to get a compromise of its value with sufficiently high values of two other measures - specificity and total accuracy.

First we can observe that for all data sets the new filtering approach improved the sensitivity of rule classifiers comparing to the standard classifier. For some data sets the increases were quite high, see e.g. breast cancer Slovenia - 0.32, ecoli - 0.358, pima - 0.3. Considering this criterion and other approaches the new filtering is generally better than simple random under-sampling and over-sampling. The only exception is glass, where under-sampling was the first. We could also say that for other data sets under-sampling usually led to higher sensitivity than over-sampling.

However, the improvement of sensitivity may be associated with the decrease of the specificity measure – see e.g. results for bupa or pima. On the other hand considering both measures added together we can conclude that the gain of introducing the new selective filtering is still the highest – even for these worse data.

The similar observation concerns decreasing the total classification accuracy while improving the sensitivity. However, for the majority of data set this decrease may be accepted. Here, we can notice the random over-sampling is the most robust and maintains the accuracy.

To sum up, these experimental results show that the new introduced approach, which contains a selective filtering phase before inducing rule, leads to improving the sensitivity of rule classifiers and it is competitive to popular sampling techniques.

Finally, we remark that constructing classifiers from imbalanced data requires special extensions and it is still an open research field, where other concepts of changing phase of inducing classifiers are possible besides pre-processing, e.g. by changing too greedy search strategies or classification policy. Such more advanced methods are the subject of ongoing research.

6. References

[1] Batista G., Prati R., Monard M.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletters, Vol. 6 no. 1, 2004, 20-29.

[2] Blake C., Koegh E., Mertz C.J.: Repository of Machine Learning, University of California at Irvine. See the WWW link [http://www.ics.uci.edu/~mlearn/MLRepositoru.html].

[3] Chawla N., Bowyer K., Hall L., Kegelmeyer W.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Re-search, vol. 16, 2002, 341-378.

[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. In: Proc. of the AAAI Workshop Learning from Imbalanced Data Sets at the 17th Conference on AI, AAAI-2000, Austin, July 30-31, 2000, 69-74.

[5] Grzymala-Busse J.W., Stefanowski J. Wilk Sz: A comparison of two ap-proaches to data mining from imbalanced data. In: Proc. of the KES 2004 - 8-th Int. Conf. on Knowledge-based Intelligent Information & Engineering Systems, Springer LNCS vol. 3213, 2004, 757-763.

[6] Kubat M., Matwin S.: Addressing the curse of imbalanced training sets: one-side selection. In: Proc. of 14th Int. Conf. on Machine Learning, 1997, 179-186.

[7] Laurikkala J.: Improving identification of difficult small classes by balancing class distribution. Technical Report A-2001-2, University of Tampere, 2001.

[8] Stefanowski J.: The rough set based rule induction technique for classification problems. In: Proc. of 6th European Conf. on Intelligent Techniques and Soft Computing EUFIT'98, Aachen 7-10 Sept. 1998, 109-113.

[9] Stefanowski J.: Algorithms of rule induction for knowledge discovery (In Polish). Habilitation Thesis published as Series "Rozprawy", no. 361, Poznań, University of Technology Press, Poznań, 2001.

[10] Stefanowski J., Wilk Sz.: Combining rough sets and rule based classifiers for handling imbalanced data. Proceedings of the XIV Concurrency Speci-fication & Programming Conference - CS&P 2005, vol. 2. 2005, 497-508.

[11] Weiss G.: Mining with rarity: a unifying framework. ACM SIGKDD Explorations Newsletter, vol. 6 no. 1, 2004, 7-19.

[12] Wilson D.R., Martinez T., Reduction techniques for instance-based learning algorithms. Machine Learning Journal, vol. 38, 2000, 257-286.

Artykul recenzowany