Classification with Class Overlapping: A Systematic Study

Haitao Xiong[1] **Junjie Wu**[1] **Lu Liu**[1]

[1]School of Economics and Management, Beihang University, Beijing 100191, China

Abstract

Class overlapping has long been regarded as one of the toughest pervasive problems in classification. When it is combined with class imbalance problem, the situation becomes even more complicated, with few works in the literature addressing this combinative effect. In this paper, we pay a systematic study on the class overlapping problem and its interrelationship with the class imbalance problem. Five widely used classifiers and three overlapping-class modeling schemes are employed for the comparative study. Extensive experimental studies on various real-world data sets reveal that: (1) The separating scheme is the best among the three schemes; (2) The distance-based classifiers are more sensitive than the rule-based ones to the class overlapping problem; (3) As the increase of the class imbalance ratio, the separating scheme showed higher improvements to the classification performance, in particular for the well-known SVMs.

Keywords: Keywords-Data Mining, Classification; Class Overlapping, Class Imbalance, Support Vector Data Description (SVDD)

1. Introduction

When dealing with classification tasks such as fraud detection, network intrusion detection, and character recognition, there is often the case that some samples from different classes have very similar characteristics. These samples are called overlapping samples for they usually reside in overlapping regions in the feature space. They also cause the so-called "class overlapping problem", which has become one of the toughest problems in machine learning and data mining communities. Indeed, researchers have found that misclassification often occurs near class boundaries where overlapping usually occurs as well [1], and it is hard to find a feasible solution for it [2].

Given its importance and difficulty, in the literature, great research efforts have been dedicated to the class overlapping problem. Prati et al. [3] developed a systematic study using a set of artificially generated datasets. Results showed that the degree of class overlapping had a strong correlation with class imbalance. These researches, however, mainly work on artificially generated datasets and focus on the effectiveness of basic classifiers in the presence of class overlapping. Therefore, from a practice point of view, there is still a critical need in conducting a systematic study on the schemes that can find and handle overlapping regions for the real-world data sets.

Our work in this paper aims to fill this crucial void. Specifically, we have three major contributions as fol-

lows. First, we use the one-class classification algorithm Support Vector Data Description (SVDD) [4] creatively to capture the overlapping regions in real-world data sets. Second, we present three simple schemes, namely the *discarding*, *merging* and *separating* schemes, to model the data sets with the presence of class overlapping. These schemes were mainly borrowed from the handwriting recognition [1, 5] and multi-label classification [6, 7] studies, although they found overlapping samples by the given labels only. Finally, we study the interrelationships between the best *separating* scheme and five popular classifiers, with and without the impact of class imbalance [8, 9].

2. Methodology

2.1. Schemes for finding overlapping regions

The objective of Support Vector Data Description (SVDD) is to find a sphere or domain with minimum volume containing all or most of the data [4]. Let$X$=$\{x_i,\ \ i$=$1 \ldots l\}$ be the given training data set of l points. In order to find the domain with minimum volume, we should minimize the radius of sphere with the constraint that the distance between a center and data is smaller than the radius. Suppose that a and$R$ denote center and radius respectively, the following function is minimized:

$$\begin{aligned} \min_{R,\xi} \quad & q(R,\xi) = R^2 + C\sum_{i=1}^{l}\xi_i \\ \text{s.t.} \quad & ||\Phi(x_i) - a||^2 \le R^2 + \xi_i, \\ & \xi_i \ge 0, i = 1 \ldots l. \end{aligned} \tag{1}$$

where the parameter C gives the tradeoff between the volume of sphere and the number of error, and slack variables ξ incorporates soft constrains. To solve the Eq. (1), the Lagrangian is introduced. For minimizing the Lagrangian function L, setting the derivative of L to zero with respect to R, a and ξ, respectively, leads to:

$$\begin{aligned} \sum_{i=1}^{l}\alpha_i = 1, \\ a = \sum_{i=1}^{l}\alpha_i\Phi(x_i), \\ \alpha_i = \tfrac{1}{l\nu} - \beta_i. \end{aligned} \tag{2}$$

According to the Karush-Kuhn-Tucker (KKT) optimality conditions, we have:

$$||\Phi(x_i) - a||^2 < R^2 \Rightarrow \alpha_i = 0. \tag{3}$$

$$||\Phi(x_i) - a||^2 = R^2 \Rightarrow 0 < \alpha_i < C. \tag{4}$$

$$||\Phi(x_i) - a||^2 > R^2 \Rightarrow \alpha_i = C. \tag{5}$$

From the above results, the domain description is represented in terms of only support vectors whose Lagrange multipliers satisfy $0 < \alpha_i < C$. Support vectors define the boundary of training data. So, SVDD can be used to find overlapping region, as shown in Fig. 1. Red dotted and blue dashed line are the boundaries of the two classes. The data dropped in both two spheres can be thought as the overlapping data which is close to or overlaps with each other.

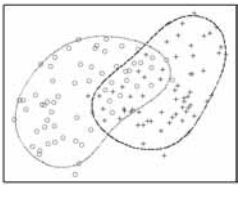

Fig. 1: Boundaries described by SVDD.

2.2. Schemes for dealing with class overlapping problem

The research on this topic mainly focuses on schemes for dealing with class

overlapping problem, which can be divided into two strategies. One is to merge the overlapping classes into a meta-class and ignore the boundary between these classes. The other is to separate the overlapping classes by refining the classification boundary in the feature space.

In this paper, we propose and evaluate three different overlapping-class modeling schemes, namely the *discarding*, *merging* and *separating* schemes, to handle class overlapping problem for the comparative study.

The *discarding* scheme (*d*): this scheme ignores the data in overlapping regions and learns on the data in non-overlapping region. Each test data will be tested on the model learned by this scheme.

The *merging* scheme (*m*): this scheme merges the data from overlapping regions as a new class, labeled 'overlapping'. Then the scheme focus on the whole data with 'overlapping' class added. After that, the scheme continues to learn on the data in overlapping region. So, two models can be got. One is the model of the data that adds 'overlapping' class, and the other is the model of the data in overlapping region. Each test data will be firstly tested on the first model. If it is classified as 'overlapping', it will be tested on the second model to determine its original class.

The *separating* scheme (*s*): this scheme learns on data in overlapping and non-overlapping regions separately. It can get two models. One is the model of the data in overlapping region, and the other is the model of the data in non-overlapping region. Each test data will be tested on the model which is determined by that if the test data is lying on non-overlapping regions or overlapping regions.

3. Experimental Results

The main goal of this research is to gain some insights on class overlapping problem and its interrelationship with class imbalance problem. We will investigate how schemes dealing with class overlapping problem affect the performance of learning in such conditions.

To make this comparison, five real-world binary data sets from UCI and LIBSVM which have different overlapping ratios calculated by SVDD are selected. Table 1 summarizes the data sets. For each data set, it shows the number of examples (#E), number of attributes (#A) and the ratio of overlapping(ratio). Experiments are divided into three scenarios. The first one uses three different schemes and five distinct classifiers to deal with class overlapping problem for data sets with different overlapping ratios. The second one compares the original scheme with the best scheme in first experiment and finds what may affect classification performance. The third one concentrates on imbalance data sets generated by random sampling to find combinative effect.

Data Set	Source	#E	#A	ratio
Inosphere	UCI	351	34	41.3%
Sonar	UCI	208	60	20.2%
Splice	LIBSVM	1000	60	12.6%
Diabetes	UCI	768	8	7.9%
German	LIBSVM	1000	24	5.7%

Table 1: Characteristics of Data Sets.

The experiments adopt a ten-fold cross-validation method. F-measure metric of the both classes is used to compare the performance of five classification algorithms with distinct natures: NB, k-NN (k=5), SVMs (with linear kernel), C4.5 and RIPPER.

3.1. Experiment 1

The first experiment is over a collection of above five binary data sets with decreasing class overlapping ratios. For each data set, the overlapping regions are gained by SVDD. So each data set can be divided into two parts: non-overlapping and overlapping. After that, three overlapping-class modeling schemes described in previous section and five distinct classifiers will be used. F-measure of both classes will be calculated for comparison.

Table 2 shows the performances of the three schemes and five classifiers on five binary data sets. The values after the name of data set are the overlapping ratio of each data set and the results in bold indicate the best values. As can be seen, F-measures of the *separating* scheme for the both classes in binary data sets are always higher than the *merging* scheme. This is because the *merging* scheme introduces additional class to the data set, which makes the classification become more complicated.

Another observation is that for two data sets, *Inosphere* and *Sonar* with overlapping ratio more than 20%, the *separating* scheme performs almost better than the *discarding* scheme. While for the remaining three data sets with overlapping ratio less than 20%, performance of the *separating* scheme is close to the *discarding* scheme. As the increase of the value of overlapping ratio, the *separating* scheme performs much better than the *discarding* scheme and the performance gap is broad. This is because ignoring this data may lose some important information of each two classes. If the overlapping ratio is high, more information will be discarded to bring lower F-measure.

These observations support that the *separating* scheme which learns on overlapping and non-overlapping data separately is the best among the three schemes. The data sets used in the *separating* scheme are divided into small part which will be good for precisely learning and time costing.

3.2. Experiment 2

The *separating* scheme that deals with class overlapping problem performs best is also compared to the scheme without dealing with class overlapping problem called the *original* scheme (o) here. The second experiment is designed to find out what may affect the performance of overlapping-class modeling schemes. Data sets and classifiers used in this experiment are the same as those in pervious experiment.

Fig. 2, in which F1 and F2 means F-measure for class 1 and class 2, shows the comparing results of the *separating* and *original* schemes. Data sets from left to right are sorted by the overlapping ratio in ascending order. One observation is that, for all data sets and classifiers, the *separating* scheme performs better than the *original* scheme or has almost the same results as the *original* scheme. Another observation is that as the overlapping ratio increases, the gap between the *separating* and *original* schemes is more and more obvious except for those high F-measure values. This is reasonable for that high F-measure can difficultly be improved. And because low overlapping ratio means that there are little data lying on overlapping region, overlapping-class modeling schemes can hardly improve the overall performance.

In addition, we also investigate how the classifiers affect the performance.

Data sets	F-measure(class 1)			F-measure(class 2)		
	d	*m*	*s*	*d*	*m*	*s*
Inosphere						
C4.5	0.766	0.809	**0.863**	0.788	0.902	**0.925**
NB	0.607	0.849	**0.861**	0.407	0.909	**0.914**
k-NN	0.722	0.668	**0.805**	0.797	0.873	**0.910**
SVMs	0.629	0.760	**0.786**	0.612	0.897	**0.905**
RIPPER	0.728	0.809	**0.848**	0.730	0.892	**0.918**
Sonar						
C4.5	0.620	0.582	**0.652**	0.593	0.559	**0.701**
NB	0.689	0.560	**0.715**	0.479	**0.673**	0.663
k-NN	0.642	0.615	**0.651**	0.569	0.623	**0.645**
SVMs	**0.699**	0.663	0.676	0.665	0.689	**0.719**
RIPPER	0.678	0.613	**0.678**	0.576	0.512	**0.640**
Splice						
C4.5	**0.936**	0.926	0.928	**0.939**	0.929	0.930
NB	0.821	0.826	**0.829**	0.833	0.831	**0.837**
k-NN	0.775	0.726	**0.784**	0.723	0.551	**0.726**
SVMs	0.787	**0.790**	0.783	0.803	**0.804**	0.794
RIPPER	0.936	0.878	**0.941**	0.939	0.880	**0.943**
Diabetes						
C4.5	**0.622**	0.604	0.606	**0.804**	0.797	0.795
NB	0.623	0.620	**0.625**	0.8236	0.817	**0.832**
k-NN	0.608	0.600	**0.611**	0.807	0.796	**0.809**
SVMs	**0.629**	0.617	**0.629**	**0.832**	0.824	**0.832**
RIPPER	**0.632**	0.625	0.631	0.816	0.802	**0.816**
German						
C4.5	0.523	0.469	**0.537**	**0.828**	0.795	0.822
NB	0.548	0.553	**0.557**	**0.835**	0.808	0.830
k-NN	0.409	0.396	**0.423**	0.797	0.798	**0.801**
SVMs	0.514	0.514	**0.541**	**0.836**	**0.836**	0.832
RIPPER	0.483	**0.536**	0.521	0.823	0.829	**0.830**

Table 2: Classification Results by Three Schemes and Five Classifiers.

Five classifiers can be divided into two types. C4.5 and RIPPER are rule-based classifiers. NB, *k*-NN and SVMs are distance-based classifiers. The C4.5 algorithm employs a greedy technique to induce decision trees for classification which is a locally classification algorithm. So the scheme for dealing with class overlapping problem can hardly work well in C4.5. For other four classifier (NB, *k*-NN, SVMs and RIPPER), scheme for dealing with class overlapping problem is useful for classification. But there is a special one which is RIPPER. As can also be seen, using the *separating* scheme, RIPPER performs better on data set *Diabetes* than the *original* scheme and performs the same on data set *Sonar*. These are converse to the results of NB, *k*-NN and SVMs. The reason is that RIPPER builds rules first for the smallest class and will not build rules for the largest class. But the largest class in overall regions may be not the largest class in overlapping regions.

In summary, the distance-based classifiers are more sensitive than the rule-based ones to the class overlapping problem.

3.3. Experiment 3

Following previous experiment, this experiment concentrates on three distance-based classifiers (NB, *k*-NN and SVMs) for imbalance data set. In order to generate imbalance data set, the data set with overlapping ratio more than 10% is selected and random sampling on the small class is used. The sampling ratio was set to 0.2, 0.4, 0.6, 0.8.

Fig. 3 shows the classification results

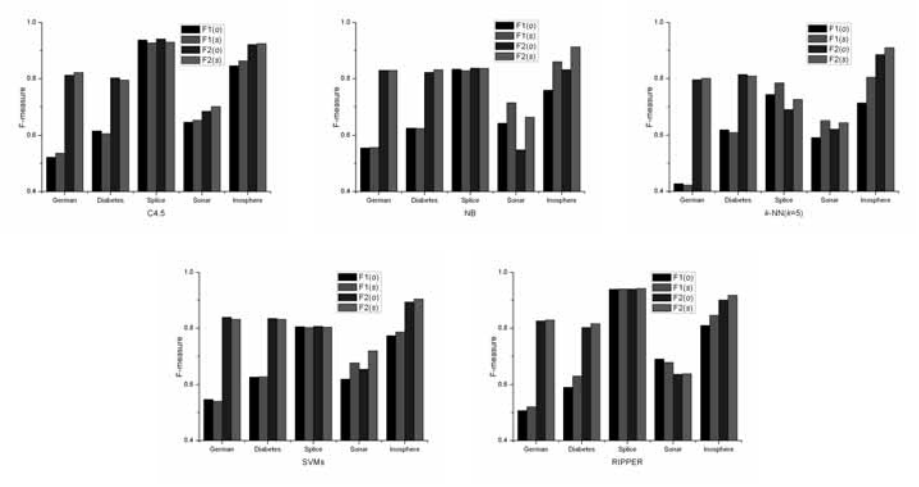

Fig. 2: Performances of the *separating* and *original* schemes.

of the two schemes on these samples. As can be seen, on data sets which have high overlapping ratios, the F-measure values of the *separating* scheme are consistently higher than the *original* scheme, no matter what the sampling ratios are. Moreover, the *separating* scheme performs much better than the *original* scheme when the rare class size is relatively small and the overlapping ratio is relatively high. As the increase of the size of the rare class, the performance gap for SVMs is narrowed. However, for NB and k-NN, the performance gap is not obviously narrowed. In summary, as the increase of the degree of class imbalance, the *separating* scheme showed higher improvements to the classification performance, in particular for the well-known SVMs.

4. Conclusion

In this paper, we took a systematic study on the modeling schemes that were proposed specifically for handling class overlapping problem. SVDD was first employed to find overlapping regions in the data. We found that modeling the overlapping and non-overlapping regions separately was the best scheme for solving the class overlapping problem. By further comparative studies on five widely used classifiers, we found that the distance-based classifiers such as SVMs and k-NN were more sensitive than the rule-based ones to the class overlapping problem, and therefore enjoyed a much better performance using the *separating* scheme. Finally, when combined with the class imbalance problem, we found that as the increase of the degree of imbalance, the *separating* scheme showed higher improvements to the classification performance, in particular for SVMs.

Acknowledgements

The research was supported by the National Natural Science Foundation of China under Grant Nos. 70901002, 90924020, and the PhD Program Foundation of Education Ministry of China under Contract Nos. 200800060005, 20091102120014.

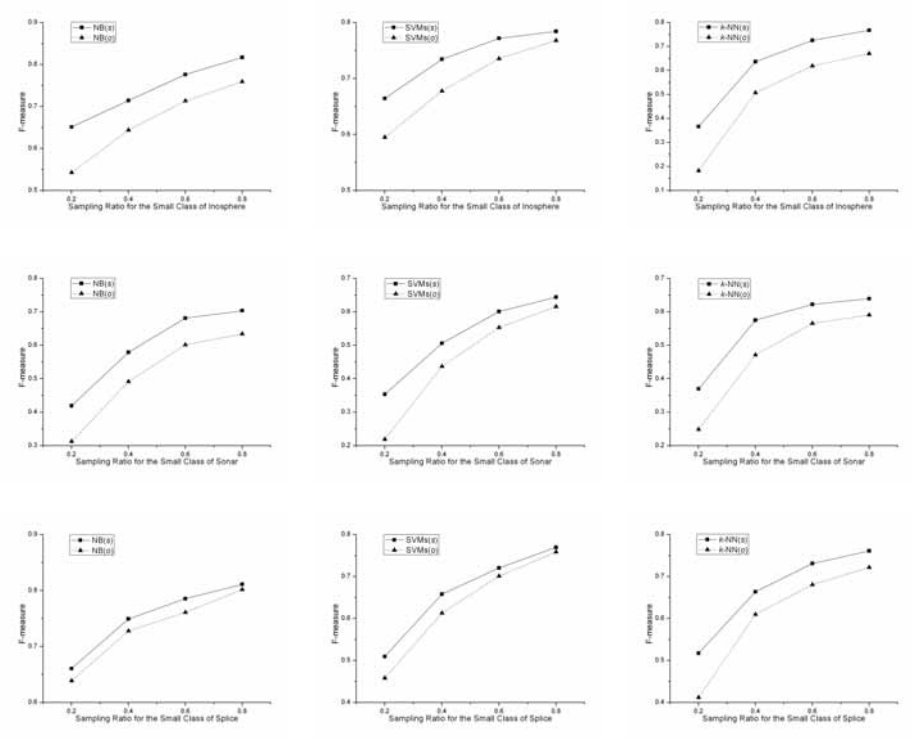

Fig. 3: The effect of class imbalance on the *separating* and *original* schemes.

References

[1] C.L. Liu. Partial discriminative training for classification of overlapping classes in document analysis. *IJOC*, 11(2):53–65, 2008.

[2] V. García, R. Alejo, J. S. Sánchez, J. M. Sotoca, and R. A. Mollineda. Combined effects of class imbalance and class overlap on instance-based classification. In *IDEAL*, pages 371–378, 2006.

[3] R.C. Prati, Batista G.E., and M.C. Monard. Class imbalance versus class overlapping: an analysis of a learning system behavior. In *MICAI*, pages 312–321, 2005.

[4] D. Tax and R. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, 2004.

[5] I.T. Podolak. Hierarchical classifier with overlapping class groups. *Expert Syst. Appl.*, 34(1):673–682, 2008.

[6] M.R. Boutell, J. Luo, X. Shen, and C.M. Browm. Learning multilabel scene classification. *Pattern Recognit.*, 37(9):1757–1771, 2004.

[7] G. Tsoumakas and I. Katakis. Multi–label classification: an overview. *Int. J. Data Warehousing Min.*, 3(3):1–13, 2007.

[8] T. Jo and N. Japkowicz. Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter*, 6(1):40–49, 2004.

[9] J. Wu, H. Xiong, and J. Chen. Cog: local decomposition for rare class analysis. *DMKD*, 20(2):1384–5810, 2010.