Using SVM for Classification in Datasets with Ambiguous Data

Saeed Hashemi and Thomas Trappenberg
Faculty of Computer Science
Dalhousie University
6050 University Ave.
Halifax, B3H 1W5, Canada
(saeed,tt)@cs.dal.ca

ABSTRACT:

One of the challenges in machine learning is the classification of datasets with ambiguous instances. In this paper we study specifically datasets with examples that have overlapping feature values for different classes. In these circumstances there is a bound on the classification performance. While there seems to be a race for accuracy, very little has been done to understand and solve the issues related to ambiguous data where the possible classification performance is limited. We discuss the use of SVMs in a proposed scheme to handle classification in such problem domains. A new approach is offered that tries to separate ambiguous data from the data that are much simpler to classify in order to prevent their influence on the classification process. We demonstrate that by separating the ambiguous data, although we lose some data, the performance of the classification increases significantly. In contrast to previous findings with some other classifiers, our experimental results show that the performance of SVM classifiers on cleaned data is not affected significantly when there are some atypical points in the training data.

Keywords: ambiguous data, atypical, outliers, overlapping samples, SVM, classification.

1. INTRODUCTION

Atypical data is often a source of concern in any classification process. Atypicality can show itself in different ways. Outliers, as one kind of atypical data, have attracted researchers' attention for a long time. Ripley [1] defines outliers as "examples which did not (or thought not to have) come from the assume population of examples." Barnett and Lewis [2] have almost the same definition for outliers: "an observation (or subset of observations) which appear to be inconsistent with the remainder of the set of data." Most definitions of outliers specify that such examples raise the suspicion that they are from a different distribution than the rest of the dataset.

Ambiguous data due to overlapping feature values, as we define here, is a different type of atypicality. Unlike outliers, ambiguous data do not show any inconsistency with the other datapoints in the same class. Thus, techniques like residual analysis and different distance measures like Cook's distance [3] cannot distinguish them.

One specific example of ambiguous data with overlapping samples that cause a serious problem to the classification task is shown in Figure 1. Figure 1A shows a training dataset with two attributes and 50 datapoints for each of two classes. The first attribute, x_1, varies uniformly within the interval [-0.8, 0.2] for class 1 and [-.02, 0.8] for class 2. The second attribute, x_2, is also uniformly distributed within [0, 1] and is included only to help to demonstrate the data. Due to the overlapping attribute values in x_1, there is no way to train an algorithm to classify the datapoints in the region x_1 = [-0.2, 0.2] because data points in this region have equal probability to belong to either of the two classes. Thus, even if an algorithm produces no error in the training set, the upper bound in classification performance in this example is only 80%; 100% in the non-overlapping regions, and 50% in the overlapping region.

Figure 1B shows the result of training a support vector machine (SVM) applying a RBF kernel function; all training examples are correctly classified. Figure 1C shows the result of applying the trained SVM on test data; only 80% of the data were classified correctly.

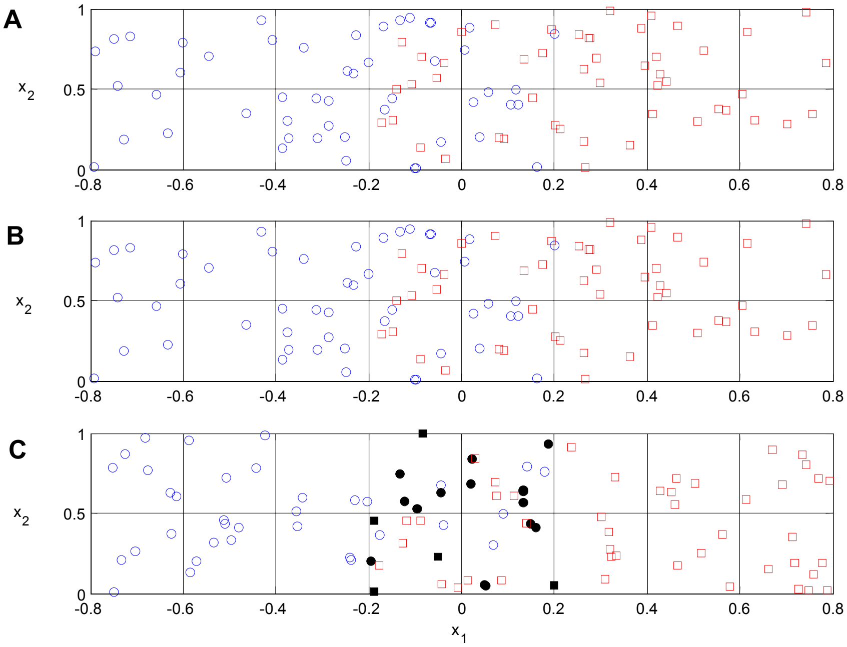

Figure 1. Example of uniformly distributed overlapping samples. (A): training data; circles are class 1 and squares are class 2. (B): The classification of training data. (C): The classification of test data; false classifications are marked with solid symbols.

Different schemes have been suggested to work with atypical data. One approach is to perform an unsupervised learning on them. This includes clustering methods as a preprocessing technique and 1-class SVM [4]. The other approach is to use a hybrid system [5] to detect and classify them. Trappenberg and Back [6] have suggested the idea of adding a new class IDK (I Do not Know) to the number of target labels, and to classify atypical points into the IDK class.

The scheme we propose here is different from the previous ones in that the atypical points are tried to be kept separate from the rest of the dataset. Since atypical points are often highly influential, keeping them within the same dataset may cause the misclassification of some of the regular datapoints. This was motivated by the previous findings [6] that the performance of a separation scheme is better than a straight forward classification due to reducing the effect of the presence of atypical examples on the classification process. This scheme is explained in section 2, the implementation issues, the results of our experiments, and a discussion are presented in sections 3, 4 and 5 respectively, before concluding in section 6.

2. SEPARATION SCHEME

The following scheme explains our approach. The scheme is general enough to work for both outliers and ambiguous data but we only target ambiguous data here.

On training data:

1. Train classifier 1 using all the training data.
2. Use the information from classifier 1 to divide all points into 2 classes: A (typical) and B (atypical).
3. Train classifier 2 on the training data with new labels (A and B); classifier 2 is atypical detector (separator).
4. Train an additional classifier 3 on only A (typical data) using their original labels.

On test data:

5. Use classifier 2 to remove potential atypical data from the test set (cleaning test data): 2 classes A1 and B1.
6. Use classifier 3 for the classification of A1 data. Use original labels to calculate the performance measures.

In the second step above, we use some measure to separate the ambiguous datapoints. This can be done by, for instance, assigning a threshold on posterior probability in probabilistic classifiers, the number of same-class datapoints found in a KNN algorithm [6], or choosing the bounded support vectors (BSVs) in the case of a SVM [7]. The function of classifier 2 is to separate the ambiguous data from the typical ones. In the experiments reported below we calculate in addition to the performance of classifier 3 (SVM3 in this study) in step 6, the performance of classifier 1 (SVM1 in this study) in the same step in order to compare the performance of these two classifiers. We expect that the performance of SVM3 is considerably higher than SVM1.

We calculate a curve that shows the coverage versus performance (CP curve) to find out how many and which datapoints to take away form a dataset to have a better classification on the clean (typical) data. In general, a CP curve is calculated by first taking away some minimum number of atypical examples in step 2, finishing through step 6, and repeating this process from step 1 to take away some more potentially atypical points from the training set. The reason for such a gradual approach is that (1) atypical datapoints are usually influential and training should be done for any new subset of training data; and (2) we do not know, in advance, which points are atypical. Coverage is calculated from the test data as coverage = (number of examples in class A1) / (total number of examples in test data).

3. IMPLEMENTATION

Every classification algorithm that can somehow distinguish atypical data from regular data can be used in the above scheme. SVMs attracted much attention in recent years particularly in the area of classification [8] and novelty detection [4]. A SVM was used in this study because of its generally high performance, and the possibility that it can be tuned to generate different number of BSVs within a dataset by changing the regularization parameter C. In the case of a ν–SVM, one can change the value of ν in every iteration of the above scheme to generate a CP curve. ν is an upper bound on the fraction of outliers and a lower bound on the fraction of support vectors [9].

An RBF kernel function was used within the SVM. To obtain a good performance, two parameters have to be chosen carefully. These parameters include C and σ. C determines the tradeoff between training error and minimizing model complexity; and parameter σ of the RBF function defines the nonlinear mapping from input space to some high dimensional feature space.

We take a fixed σ (obtained initially by parameter optimization) and apply different C values. Each C value gives a different number of BSVs on the train data. BSVs are the most qualified candidates for being atypical datapoints if their number is chosen properly. This is because they have the largest Lagrange multipliers [7]. Note that the number of atypical points is often unknown and that a CP curve can be used to estimate it.

Each time a new C is chosen, we start from step 1 (training with all training data). Thus, points on the CP curve are independent of each other. We found that the resulting coverage by varying C is very sensitive to the C value, leading to clusters with examples around large and small coverage values. In practice we repeated the experiments with different datasets many times so that a sufficient number of examples for intermediary coverage values were found. A better tuning to intermediary coverage values might be obtained with ν–SVMs. This issue is worth studying in more detail in future work.

4. EXPERIMENTAL RESULTS

Two generated datasets were used in the experiments. One was derived from uniformly distributed data as described in section 1. The second dataset was derived from normally distributed data (Gaussian). For Gaussian data, x_1 has the variance of $\sigma^2 = 1$ and the mean values of -1 and 1 for class 1 and class 2, respectively. The second attribute, x_2, is also uniformly distributed within [0, 1].

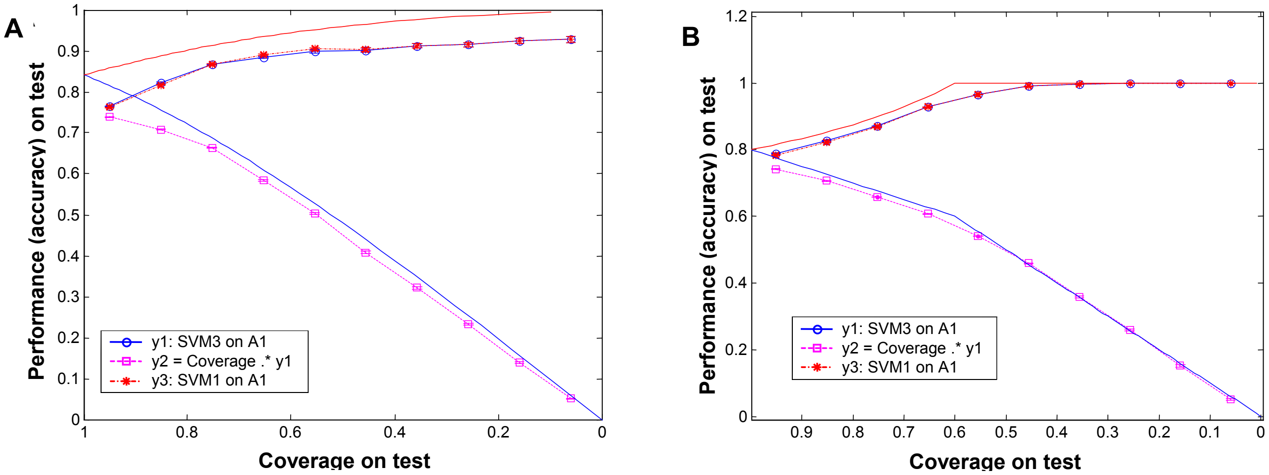

Figure 2. Coverage performance (CP) curves. Curves without markers are theoretical limits and the ones with markers are experimental results for normally distributed data (A) and uniformly distributed data (B).

In Figure 2, there are two sets of CP curves representing the results from the datasets with normally distributed (A) and uniformly distributed data (B). In each of these results, the set of monotonically increasing curves represent the performance as measured by the number of correct classifications relative to the number of classified examples (clean data). In contrast, the decreasing curves represent the performance as measured by the number of correct classifications relative to the number of all examples, including non-classified examples. In other words, the non-classified data are simply considered as misclassifications in this performance measure. Other choices of performance measures, which give different weights to the accuracy and non-classification of data, are possible and depend on the specific application [10]. The curves above thus represent the bounds on any reasonable performance measure.

The solid lines without any marker on them represent the theoretical limits of the performance measures, which can be calculated analytically for these examples. These theoretical limits were calculated considering the known distributions of the generated datasets and applying the rule that the class with the largest posterior probability, provided that it is larger than some threshold, is chosen as the predicted class. Examples without a posterior probability larger than the set threshold value are considered as ambiguous. Thus, by changing the threshold value we get different coverage values. The performance for these coverage values can be calculated from the posterior probabilities of the classified data.

The results of the SVM classifications are shown with different symbols that are interpolated by lines. y1 and y2 represents the results of SVM3. y1 (circles) is the performance when taking only the number of clean datapoints into account, whereas y2 (squares) represents the performance evaluated relative to all datapoints. y2 can also be calculated by y2 = (coverage.* y1).

y3 (asterisks) represents the performance of SVM1 when taking only the clean data in the performance measure into account. These datapoints are very close to the datapoints of curve y1 representing the corresponding results of SVM3 and are thus difficult to distinguish in this plot.

Errorbars in the experimental results (both, in Figures 2 and in Figure 3 below) are calculated as the standard deviation from 100 different datasets generated randomly. The errorbars in Figure 2 are in the order of the symbol sizes. The different coverage values are binned into a fixed number of bins, and the performance measures are averaged within each bin. This average value of performance is assigned to the midpoint of any bin to represent a point on the CP curve. Note that the performance value of the midpoint coverage is not necessarily equal to the calculated average. We call this difference the binning error.

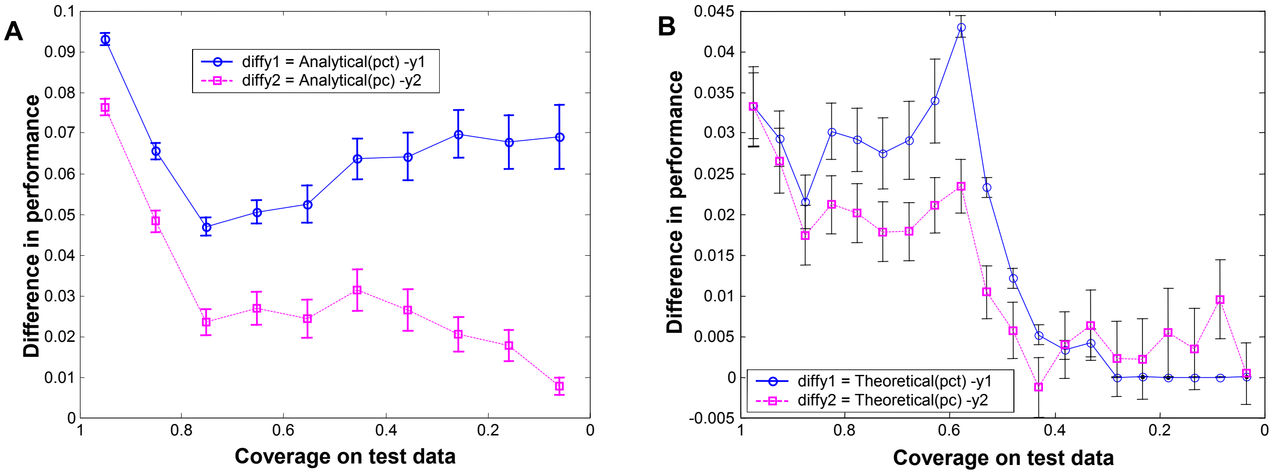

Figure 3. Difference curves (A): Normally distributed data (B): Uniformly distributed data

5. DISCUSSION

The results demonstrate that the performance of the classifiers does depend strongly on the ambiguous data. When we evaluate the performance of the classifiers only with respect to the cleaned data (rising curves), which is appropriate if accuracy rather than completeness is important in an application, we find that separating ambiguous data can largely enhance the performance of the classification process.

A further result is that the SVM classifier can achieve performances closer to the theoretical limit on cleaned datasets. To better illustrate this point we show in Figure 3 the difference in performance between theoretical limits and the performance of the SVM classifier as a function of the coverage. This indicates that the relative performance becomes better by separating potential atypical points. In particular, even removing only a small fraction of ambiguous data points can considerably enhance the classification ability of the SVM classifier. Practically, small coverage values are not of any interest and the results from low coverage points may be suspicious for not having sufficient data for training. The kink (sharp raise close to coverage = 0.6) in Figure 3B is partially due to the binning error, because the uniformly distributed data have a sharp boundary around these values. This can also be seen in Figure 2B.

A surprising result is that the performance of SVM3 and SVM1 in classifying the cleaned data is very similar. The difference between these two is almost undistinguishable. This is rather counter-intuition because one would usually assume removing atypical points from the training set should result in training a better classifier for the typical points. This was not the case for SVM, and is in marked contrast to other classifiers, such as the neural network classifiers as reported in [6]. Note that this does not mean that separating atypical points does not matter because both SVM3 and SVM1 are tested on separated typical datapoints, A1. This demonstrates a major advantage of SVM in that this type of classifiers is not easily disturbed by ambiguous data. This may come from the fact that SVMs rely mainly on the dominant contributions of some specific datapoints that determine the support vectors and that the existence of some atypical points in training a SVM may not degrade its performance on typical points. As a result of this finding we can skip step 4 in the classification scheme when using SVM as classifiers and use the classifier trained in step 1 for the predictions in step 6. This, however, may not be true for classifiers algorithms other than SVM.

6. CONCLUSIONS AND OUTLOOK

We discussed, in this paper, the classification of data with overlapping feature values, in which there are strong limits on the theoretical performance any classifier can achieve. We found that the theoretical derived curves for the coverage versus performance (CP curves) are paralleled when SVMs are used. All the benefits of removing ambiguous data from the classification set is thus paralleled by using SVMs as classifiers, such as in situations where theoretical CP curves are not available and have to be discovered by some machine learning method. Previous studies have shown that bounded support vectors (BSVs) can be used to separate outliers. Our study demonstrated that BSVs could also be used to separate atypical points from the typical ones in the case of ambiguous data due to overlapping feature values.

CP curves are very useful in practice as they can show if ambiguous data are a problem that might limit the performance of the classification process. In addition, these curves can be used to determine if the performance of the classification can be enhanced with reasonable relaxation of the coverage. We found that it is not easy to cover efficiently intermediate values of the coverage by fine-tuning the regularization parameter C to get different numbers of BSVs. Sometimes a small change in C leads to a large change in the number of BSVs. This makes it difficult in practice to apply the SVM to derive CP curves. We think that ν–SVM might be more efficient to derive CP curves, and further studies should investigate this.

We are currently exploring the performance of our scheme on real-world datasets, and are comparing the utilization of SVM with other classifiers in the more general separation scheme outlined in section 2. Note also that the removed points can be used in further analysis such as investigating missing attributes or even a possible extra class not considered in the training set. More research in this direction is thus desirable.

Acknowledgement

This work has used the OSU-SVM Matlab code (http://eewww.eng.ohio-state.edu/~maj/osu_svm/), which is based on LibSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm). We also appreciate Dr. Chih-Jen Lin's sincere help via email.

7. REFERENCES

[1] B.D. Ripley, “Pattern recognition and neural networks”, Cambridge Univ. Press, 1996.

[2] V. Barnett and T. Lewis, “Outliers in statistical data, 3rd ed.”, John Willey & Sons, New York, 1994.

[3] S. Weisberg S., “Applied linear regression, 2nd ed.”, John Wiley & Sons, New York, 1985.

[4] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, “Support Vector Method for Novelty Detection”, Advances in Neural Information Processing Systems 12, pp. 582-588, 2000.

[5] C. Domeniconi and D. Gunopulos, “Adaptive Nearest Neighbor Classification using Support Vector Machines”, TR, UCR-CSE-01-04, 2001.

[6] T.P. Trappenberg and A.D. Back, “A classification scheme for applications with ambiguous data”, International Joined Conference on Neural Networks, IJCNN 2000.

[7] B.E. Boser, I.M. Guyon, and V.N. Vapnik, “A Training Algorithm for Optimal Margin Classifieres”, In 5th annual workshop on computational learning theory, Pittsburgh, 1992, ACM.

[8] N. Cristianini and J. Shawe-Taylor, “An Intro. To Support Vector Machines and other kernel-based learning methods” Cambridge Univ. Press, 2000.

[9] B. Schölkopf, J.C.Platt, and A.J. Smola, “Kernel Methods for Percentile Feature Extraction”, MSR-TR-2000-22, 2000.

[10] Trappenberg T.P., Back A.D., Amari S.-I. (1999) A Performance Measure for Classification with Ambiguous Data, *BSIS Technical Reports* No.99-x, May 18, 1999.