

Bagging for One-Class Learning

David Kamm

December 13, 2008

1 Introduction

Consider the following outlier detection problem: suppose you are given an unlabeled data set and make the assumptions that one particular class is well-represented but you have no prior knowledge on how many outliers it contains. This scenario can arise in a variety of real world applications such as detecting intrusions in a network or spotting malignant tumors in medical images. Although constructing labels for the data is rarely impossible, in many cases, it may be cost prohibitive or inefficient.

Most outlier detection methods such as mixture of Gaussians and Parzen window estimators fit a distribution to the data and classify points as outliers if the density function evaluated at that point is below a certain threshold. These approaches work well if the data has a particular, known distribution. Since this is not often the case, a popular approach towards outlier detection is the one-class SVM (OC-SVM) in which unlabeled data is treated as positive examples for a particular class.

OC-SVM addresses the following problem: Given a data set drawn from an underlying probability distribution P , how do you estimate a simple subset S such that the probability a test point drawn from P lies outside of S is some a priori specified value between 0 and 1 [1]. The OC-SVM accomplishes this by mapping the data to a corresponding feature space and finding the optimal separating hyperplane between the data and the origin [1]. A more geometrically intuitive approach is Support Vector Domain Description (SVDD) which estimates the minimum hypersphere enclosing the data points in feature space. The two approaches are equivalent when the Gaussian kernel is applied, so only the OC-SVM formulation of the classifier is considered in this paper [2].

2 Motivation

The OC-SVM suffers on certain data sets since it separates outlier points off from the origin in addition to true representatives from the class [3]. To improve the OC-SVM classifier for outlier detection, this paper applies ensemble methods. The justification behind this is that, theoretically, they can boost the performance of any classifier. Because the OC-SVM captures too many outliers points in training, it can be interpreted as having an overfitting problem [3]. Bagging in particular was an attractive choice for its simplicity to implement and since it is generally thought of as a variance reduction technique.

3 Algorithm

The unlabeled input training set is denoted as $X = (x^{(i)}), i = 1 \dots m$. We make the assumption that one particular class is well-represented aside from outlier points which we have no prior knowledge on how many there are. With this assumption in mind, we can apply our one-class learning approach without worrying about the complications caused by multiple classes being represented in the data. In the OC-SVM, the training data is separated from the origin in the feature space by solving the following quadratic programming problem

$$\begin{aligned} \min_{w, \xi, \rho} & \frac{1}{2} w^T w + \frac{1}{\nu m} \sum_{i=1}^m \xi_i - \rho \\ \text{s.t.} & (w \cdot x^{(i)}) \geq \rho - \xi_i, i = 1 \dots m \\ & \xi_i \geq 0, i = 1 \dots m \end{aligned}$$

The corresponding dual problem can be altered to efficiently work in feature space by replacing the dot products between training vectors with an appropriate kernel K . The parameter ν is important since it is an upper bound on the percentage of outliers and a lower bound on the percentage of support vectors [2]. When there is prior knowledge about the number of outliers, the parameter ν can be tuned to give the optimal rejection rate, which is taken advantage of in the experiments.

Bagging works by resampling the data with replacement and training a classifier on the new data set for a pre specified number of trials. A prediction on a test point by taking the majority vote of the classifiers under a certain voting scheme. The bagging algorithm is given as follows:

1: **Bagging OC-SVM:**

2: Input: The number of classifiers N , the data set X

3: Output: A linear combination of OC-SVM classifiers $\sum_{i=1}^m C_i$

4: **for** $i = 1$ to N **do**

5: Draw m examples $(x^{(i)}, y^{(i)})$ uniformly from X with replacement to form training set D_i

6: Train a OC-SVM classifier C_i on D_i

7: Add C_i to the linear combination

8: **end for**

Aside from the free parameters regulating the OC-SVM, there are two factors that influence the performance of the bagged classifiers: the voting scheme and the number of classifiers. The more pressing decision to make is for the voting scheme. Unweighted majority vote is accepted as the standard voting scheme for bagged ensembles of classifiers and was thus implemented in the algorithm. Although there are weighted voting schemes that have been shown to perform better than unweighted majority vote on various data sets in supervised learning, it is difficult to construct such weightings on unlabeled data. The effect of varying the number of classifiers was addressed in the experiments.

4 Experiments

The algorithm was applied to two 2-D synthetic data sets where the decision boundary could be visualized easily. Two real world data sets were used to test the algorithm's overall performance

and effectiveness at detecting outliers.

4.1 Data sets

Square-noise: The true positives of the square data set are 400 points uniformly drawn from 4 strips of length 2.2 and width 0.2 which can be identified in Fig. 1. The outliers were 50 points drawn randomly from $\{(x, y) | x \in [0, 3], y \in [0, 3]\}$ and compose 1/9 of the data.

Sine-noise: The true positives of the sine data set are 500 points with $x \in [0, 2\pi]$ and $y = 0.8 * \sin(2x)$. The outlier points were 200 points drawn uniformly from the rectangle $\{(x, y) | x \in [0, 2\pi], y \in [-1.5, 1.5]\}$ and compose 2/7 of the data set.

USPS Handwritten Digits: The training set is comprised of 7291 images and the test set is comprised of 2007 images. Each image is 16x16=256 pixels and is represented by a 256 dimensional feature vector. The algorithms were trained on the 644 instances of the digit 0 in the training set and tested on all of the digits in performance evaluation experiments.

Breast Cancer: The breast cancer dataset was obtained from the UCI Machine Learning repository. The data is composed of benign and malignant samples with 10 features. A pre-scaled data set is available from the LIBSVM website, which was used in these experiments. For evaluating classifier performance, the algorithms were trained on the first 200 benign samples and tested on the remaining samples, which were composed of 244 benign samples and 239 malignant samples.

4.2 Methods

In each experiment, the radial basis formulation of the Gaussian kernel $K(x, z) = \exp(-\gamma||x - z||^2)$ was used since the data set is always separable from the origin when it is mapped into feature space by this kernel [1]. To make a comprehensive observation of the effect of varying the number of classifier, the bagging OC-SVM algorithm was ran with varying the number of classifiers N with values 10, 20, 50, 100 and 200. LIBSVM version 2.71 for MATLAB was used for the OC-SVM implementation [4]. On the real world data sets, performance of the classifier was measured by the ROC curves on the test data, which is a standard criterion for outlier detection.

In addition to the above experiments, the following outlier detection experiment was performed on both of the real world data sets. The USPS digit 0 training set was augmented with 64 instances of digits 1-9 drawn uniformly and with replacement, so the training set has 9.03 percent outliers. The testing set is composed of the same 64 outlier digits. On the breast cancer data set, 50 malignant samples were taken and added to the training set corresponding to 20 percent outliers. The test set is composed of the 50 malignant samples for the respective experiments. In these experiments, the assumption that the proportion of outliers is unknown is again dropped to make selection of the ν parameter straightforward. For each of the outlier detection experiments and the 2-D synthetic experiments, the ν parameter was set to be the fraction of outliers that were pre-specified in the construction of the training set. However, without enforcing the number of outliers the training data contains, selecting the optimal ν becomes difficult.

5 Results

Qualitatively, the bagged OC-SVMs resulted in a better fitting decision boundary around the target class as opposed to the OC-SVM in both the sine-noise and square-noise experiments. Although

the decision boundary is influenced by the kernel parameter σ for both classifiers, the bagged OC-SVM always excludes more outlier points. However, on the square noise data set decision boundary determined by the bagged OC-SVM seems to exclude a few positive points on the square as well, which was discouraging.

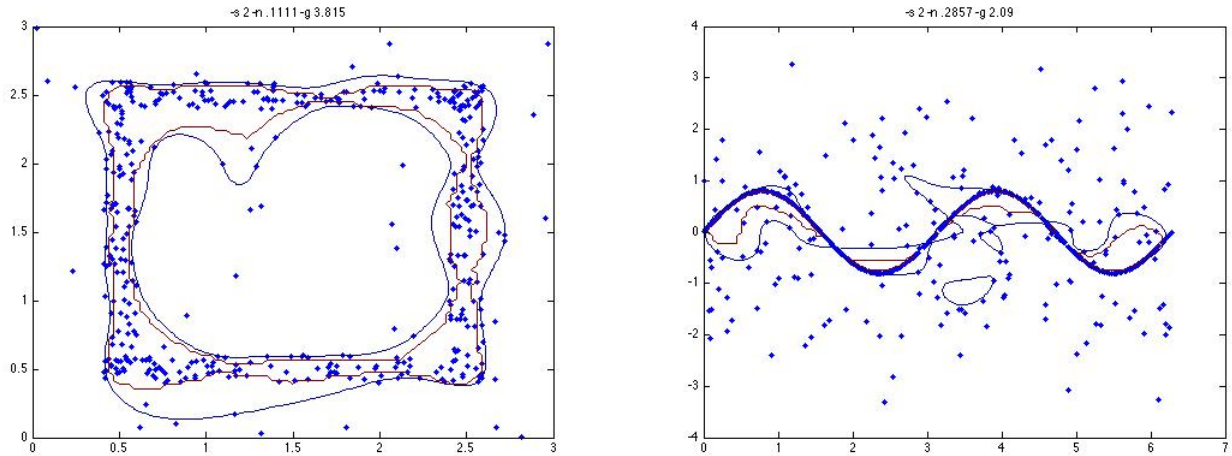


Fig. 1 The decision boundary of OC-SVM is shown in blue while the decision boundary of the bagged OC-SVMs is shown in dark red. In both cases, the bagged OC-SVMs determine a more refined boundary around the target class. In the square noise example, the kernel parameter $\gamma = 3.815$. In the sine noise example, the kernel parameter $\gamma = 2.09$. In both cases, the best number of bagged classifiers is $N = 10$.

Unfortunately, the results on the real world data sets were more discouraging. There is not a noticeable difference between the curves in the USPS data set. In the breast cancer data set, the bagged OC-SVM slightly degrades the classifier performance.

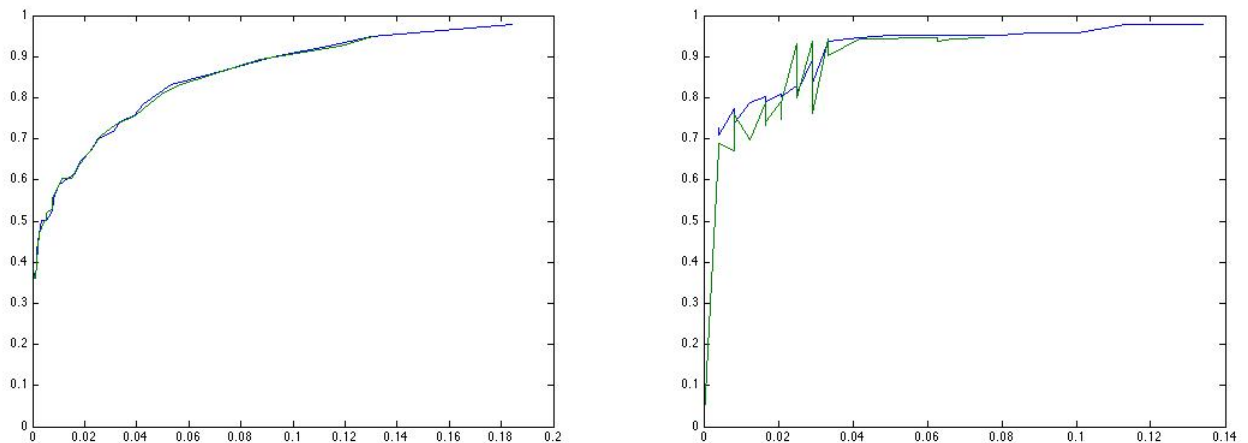


Fig. 2 Excerpts of the ROC curves for the regular and bagged OC-SVMs trained on the USPS and breast cancer data sets. In both examples, the OC-SVM curve is shown in blue and the bagged OC-SVM curve is shown in green. The kernel parameter $\gamma = 0.5$ in both experiments as well. The best number of bagged classifiers is given by $N = 10$ for the USPS data and $N = 20$ for the breast

cancer data.

The outlier detection experiments did not produce any fruitful results, either. On both data sets, the OC-SVM and bagged OC-SVM detected the same outliers in the training set when the kernel parameter γ was set to 0.5. In all of the experiments performed the optimal number of bagged classifiers was found to be either 10 or 20. Bagging over 50 classifiers always resulted in performance degradation.

6 Conclusions

Although bagging OC-SVMs improved the decision boundary on the 2-D synthetic examples, there was no noticeable improvement on the real world data sets which are much more representative of those that would be encountered in practice. Additionally, there is an N-fold increase in computational complexity when applying the bagged OC-SVM algorithm as opposed to OC-SVM. If computational requirements are an issue, any possible improvement garnered by the bagged classifier would probably be outweighed by the computational cost. Otherwise, a computationally intensive procedure such as kernel PCA as reported in [3], would be a better investment.

7 Acknowledgements

I would like to thank the TAs Tom Do and Honglak Lee for their advice. I would also like to thank the instructor Andrew Ng for teaching CS 229.

8 References

1. B. Scholkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, R. C. Williamson, Estimating the support of a high-dimensional distribution, *Neural Computation* 13.
2. D. Tax, R. Duin, Support vector domain description. *Pattern Recognition Letters*, 20:1113, 1191 1199.
3. H. Hoffman, Kernel PCA for Novelty Detection, *Pattern Recognition*
4. <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>