Supplemental Material: Multi-Class Open Set Recognition Using Probability of Inclusion

Lalit P. Jain[1], Walter J. Scheirer[1,2], and Terrance E. Boult[1,3⋆]

[1]University of Colorado Colorado Springs
[2]Harvard University
[3]Securics, Inc.

In this supplemental material we present additional plots that are meant to add to the reader's understanding of the open set recognition problem and our solution. We start with a look at SVM decision surfaces, and then offer a discussion of the difference in observed performance when Accuracy or F-measure is reported as an evaluation statistic. We also provide plots showing performance with alternate priors, and discuss some problems we encountered when attempting to apply Naive Bayes Nearest Neighbor as a comparison approach, which other researchers may be interested in.

1 SVM Decision Surfaces

In Fig. 2 of the main paper we showed the problems with existing models for two known classes ("1" and "2") when unknown classes ("?") are possible. In Fig. 1, we show that same figure again with the decision surface for the P_I-SVM. Pink is the region labeled as class 1 by just the one-class RBF SVM, light blue is the region labeled as class 1 by just the binary RBF SVM, green is the region labeled as class 1 by both the P_I-SVM and the binary RBF SVM (recall that the P_I-SVM uses a binary SVM as a basis), and magenta is the region labeled as class 1 by all three models. For the P_I-SVM, the threshold δ was set at 0.055, which is $0.5 \times \text{openness}$ for two known classes and one unknown class – the minimal open set assumption for the data. In Fig. 2, we show 3D surface plots of the decision function surfaces for all three models. Note how the P_I-SVM approaches zero away from the positive training data while the other models degrade far more slowly.

2 Multi-class Open Set Recognition Accuracy

The main paper presents the F-measure statistic for the experiments examining the binary decision component of object detection and multi-class open set recognition (we explain why in more detail below). Nonetheless, reviewers may be interested in recognition *accuracy* plots if for no other reason than to see that there are no hidden surprises. Accuracy is a statistical measure defining how well a recognition algorithm estimates correct decisions out of all decisions made. A decision can always be classified into one of four possibilities: true positives TP, true negatives TN, false positives FP, and false negatives FN. Thus accuracy is formally defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

⋆ This work was supported in part by ONR MURI N00014-08-1-0638 and NSF IIS-1320956.

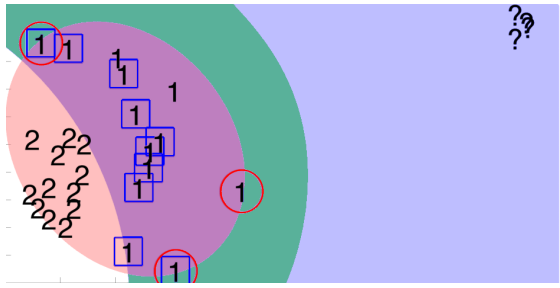
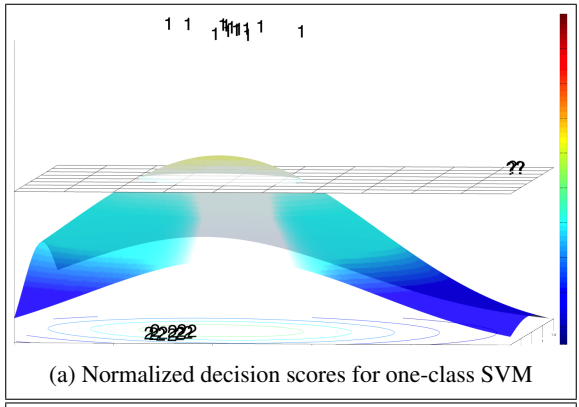


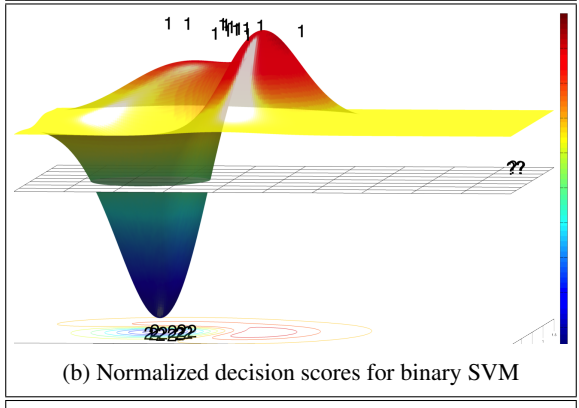
Fig. 1: The decision regions from Fig. 2 of the main paper with the addition of a P_I-SVM region. Pink is the region labeled as class 1 by just the one-class RBF SVM, light blue is the region labeled as class 1 by just the binary RBF SVM, green is the region labeled as class 1 by both the P_I-SVM and the binary RBF SVM, and magenta is the region labeled as class 1 by all three models. The one-class RBF SVM defined by the support vectors circled in red cannot separate known classes "1" and "2". The binary RBF SVM defined by the support vectors with blue squares always classifies the unknown classes ("?") as members of a known class. The P_I-SVM (also leveraging the support vectors with blue squares) does not suffer the same limitations as the other models – note the good separation between classes "1" and "2", and the right-hand decision boundary in front of the unknown classes provided by a threshold over probability of class inclusion.

Fig. 3 shows accuracy plots corresponding to the multi-class open set recognition experiments for the LETTER and MNIST benchmarks shown in Fig. 5 of the main paper. Again we see low performance for all baseline algorithms for these basic OCR tasks. Also consistent with the F-measure plots in the main paper, the P_I-SVM achieves more stability and considerably better performance over all other approaches. One curious effect in these plots, however, is the performance of the one-class RBF SVM. It is the worst among all evaluated approaches, but is also the only approach that starts improving after 0% openness (a fully closed problem). Why does this happen?

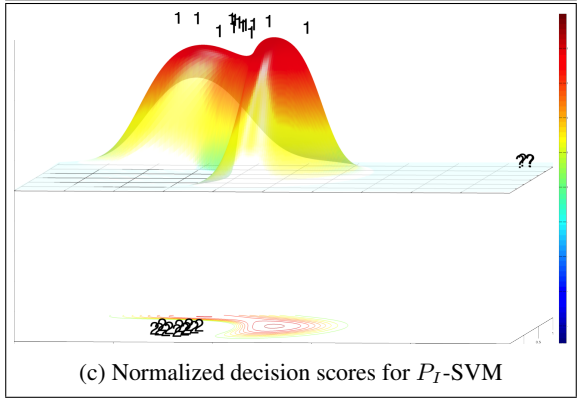
The primary reason is that the one-class SVM tends to classify most samples as negatives even if they are actually true positives. As the problem openness increases, the number of true negatives (samples from classes unseen during training) also increases, and because of its bias, the one-class SVM has an advantage in classifying such samples. Thus we see a corresponding increase in overall accuracy. The primary goal of open set recognition techniques is to predict the correct class label for a test sample if that class is known, or to reject that sample as an "unknown" if not. Since open set problems often consist of far more negative samples than positive samples, classifying most samples as negatives can actually increase recognition accuracy in evaluation. This, of course, mis-represents the actual performance of the model, which is ineffective for labeling true positives. Because recognition accuracy is not appropriate for open set problem evaluation, we would like to calculate some other statistic that gives more weight to correct positive classifications.



(a) Normalized decision scores for one-class SVM

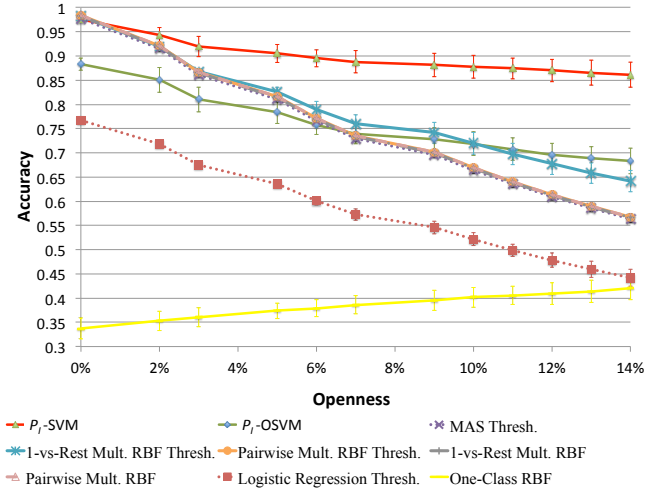


(b) Normalized decision scores for binary SVM

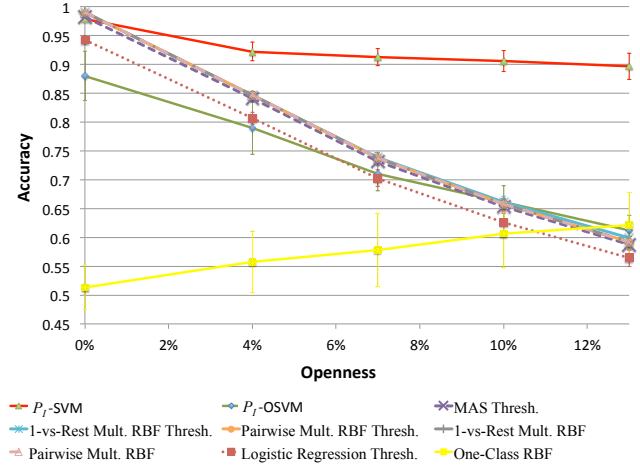


(c) Normalized decision scores for P_I-SVM

Fig. 2: Decision surfaces for (a) One-class SVM, (b) Binary SVM and (c) P_I-SVM for the example of Fig. 2 in the main paper. The black coarse mesh is at $z = 0$, with class 1 at $z = 1$, class 2 at $z = -1$ and unknowns at $z = 0$. The P_I-SVM returns to near zero away from training samples, but the binary surface stabilizes around 0.4 (far from zero) at the unknown points; the binary surface is descending so slowly that even if we extend it to [0,100]x[0,100] it is still well above zero.



(a) Multi-Class Open Set Accuracy for LETTER



(b) Multi-Class Open Set Accuracy for MNIST

Fig. 3: Multi-class open set recognition *accuracy* results for LETTER (a) and MNIST (b). These plots correspond to Fig. 5 in the main paper. Error bars reflect standard deviation, and approaches marked with "Thresh." have been augmented to support rejection. The increasing accuracy for the one-class RBF SVM highlights the problem of using accuracy as a statistical measure for recognition. Since open set problems often consist of far more negative samples than positive samples, a biased classifier that produces mostly negative decisions, such as the one-class SVM, may look better than it really is as problems grow to be more open.

3 F-measure Details

With the above accuracy result for the one-class RBF SVM in mind, we turn to a compelling argument that was raised by Scheirer et al. in Sec. 5 of [3], which advocates F-measure as a more appropriate statistic compared to accuracy for open set problem

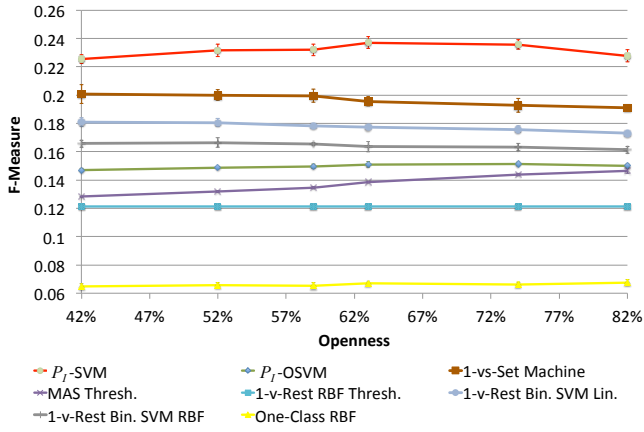


Fig. 4: F-measure performance for the binary decision component of an *open set object detection task* for an open universe of 88 classes. Results are calculated over a five-fold cross-data set style test with images from Caltech 256 used for training and images from Caltech 256 and ImageNet for testing; error bars reflect standard deviation, and approaches marked with "Thresh." have been augmented to support rejection. This figure is a replication of Fig. 4 from the main paper, with an additional curve for the one-class RBF SVM. It can be seen that the P_I-OSVM is measurably better than a standard one-class SVM (its basis), which essentially fails for this problem.

evaluation. From information retrieval, we know that precision and recall are the measurements that quantify the number of correctly classified true positive examples to all false and true positives (Eq. 2), and number of correctly classified true positive examples to all the available positive examples (Eq. 3), respectively.

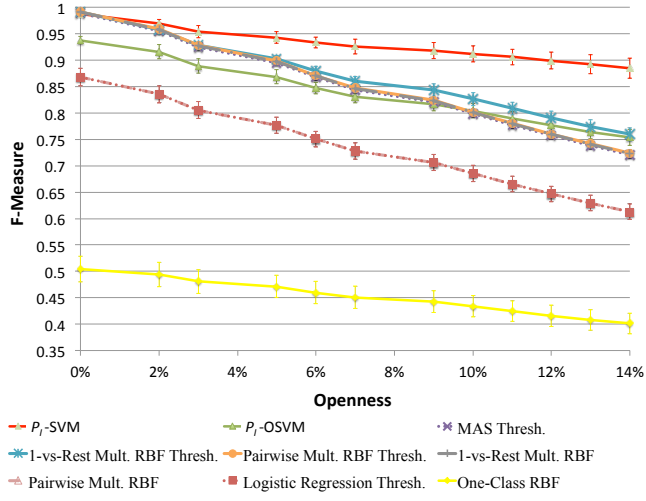
$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

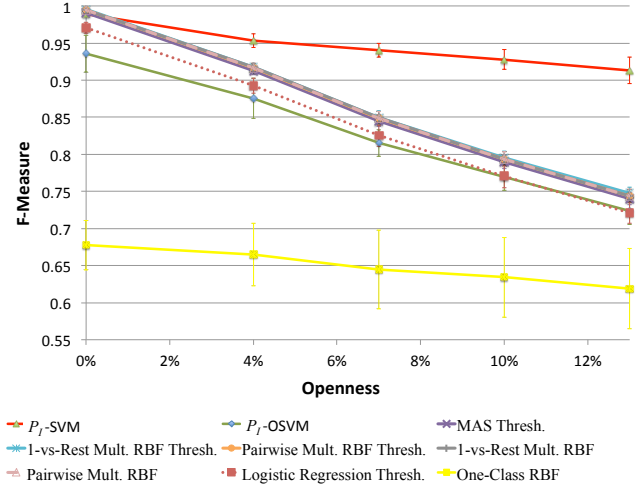
F-measure calculates a weighted average of precision and recall, allowing us to forgo calculating full Precision-Recall curves to find consistent points across each algorithm at hand. The F-measure is defined as:

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

How does F-measure compare to accuracy for open set recognition evaluation? Figs. 4 and 5 are replications of Figs. 4 and 5 in the main paper, but with the problematic one-class RBF SVM included in the plots. As F-measure more heavily weights the number of correct decisions in detecting positive classes, we no longer see improvement in terms of F-measure for the one-class SVM as openness increases. Thus misleading impressions can be avoided. The improvement observed for the P_I-OSVM over its basis, the one-class RBF SVM, can also be seen in these plots (this is not shown in the main paper).



(a) Multi-Class Open Set F-measure for LETTER



(b) Multi-Class Open Set F-measure for MNIST

Fig. 5: Multi-class open set recognition *F-measure* results for LETTER (a) and MNIST (b). These plots replicate Fig. 5 in the main paper, with additional curves for the one-class RBF SVM. Error bars reflect standard deviation, and approaches marked with "Thresh." have been augmented to support rejection. Different from the accuracy plot in Fig. 3, the one-class SVM follows the trend of all other approaches shown by decreasing in performance as the problem grows to be more open. F-measure does not inflate the performance of classifiers that have a strong negative bias. Like Fig. 4, the P_I-OSVM is substantially better than a standard one-class SVM (its basis), which performs poorly on these problems.

4 Alternate Priors

All the experiments presented in the main paper and above considered equal priors ($\rho(y)$) per class among the known classes, as described in the main paper. Reviewers might be interested to know the impact on F-measure when priors per class are computed from the known frequency of classes. This type of information is not available in all cases, but scenarios do exist where we can get good estimates. For instance, it is well known that some letters occur much more frequently (*e.g.* "e" with frequency of 12.702%) than others (*e.g.* "z" with frequency of 0.074%) in the English language. Fig. 6 shows F-measure results for the LETTER data set when setting the priors to the frequency of occurrence of letters in a natural language corpus[1]. With a significant variation in the frequency of letters in a natural language corpus, P_I-SVM retains its stability in terms of F-measure compared to other evaluated approaches.

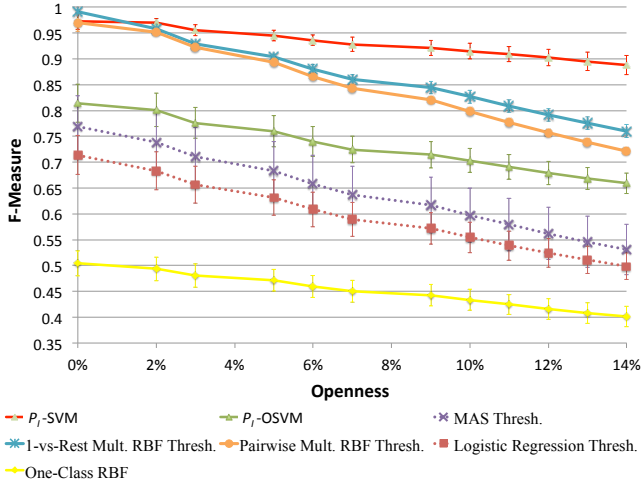


Fig. 6: The above plot shows the F-measure results for the LETTER data set when priors $\rho(y)$ in Eq. 3 of the main paper are set to the frequency of occurrence of letters in a natural language corpus. This plot corresponds to the same experiment presented in Fig. 5(a), but with only approaches producing probabilities (via calibration or inherently) as output shown. The P_I-SVM once again is more stable than existing algorithms, and achieves high F-measure scores as the level of openness increases.

5 Naive Bayes Nearest Neighbor as a Comparison Approach

While one could consider hybrid approaches such as thresholding Optimal Naive Bayes Nearest Neighbor (ONBNN) [1] for open set recognition, these algorithms do not scale very well computationally, in part because they keep all data in memory and their associated optimizations do no scale well with increasing feature dimensions. In [1], results are presented for only five classes from Caltech 101 using SIFT features from 15 down-sampled training image per class. We considered well over 1,000 times more images with 3,780 dimensional features for our object detection experiment. Using available

[1] Frequencies taken from: http://en.algoritmy.net/article/40379/Letter-frequency-English

code, we determined that for each positive class to produce one testing point required more than 72 CPU hours, compared to under a minute using our P_I-SVM. As the graph in Fig. 4 requires 5 randomized runs, 88 classes and 6 levels of openness, to complete comparative testing with ONBNN would take more than 190,000 CPU hours, or about 22 CPU years. Even with moderate parallelism, this experiment was well beyond the scope of this paper. An implementation such as [2], which uses a Local Naive Bayes Nearest Neighbor algorithm can yield up to a 100x speedup, but would only reduce the total runtime to an order of CPU months. In general, the *practical* runtime of an algorithm should always be considered in addition to theoretical optimization.

References

1. Behmo, R., Marcombes, P., , Dalalyan, A., Prinet, V.: Towards optimal naive Bayes nearest neighbor. In: ECCV (2010)
2. McCann, S., Lowe, D.: Local naive Bayes nearest neighbor for image classification. In: CVPR (2012)
3. Scheirer, W., Rocha, A., Sapkota, A., Boult, T.: Toward open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(7), 1757–1772 (2013)