Overlap versus Imbalance

Misha Denil and Thomas Trappenberg

Faculty of Computer Science
Dalhousie University
6050 University Avenue
Halifax, NS, Canada
B3H 1W5
denil@cs.dal.ca,
tt@cs.dal.ca

Abstract. In this paper we give a systematic analysis of the relationship between imbalance and overlap as factors influencing classifier performance. We demonstrate that these two factors have interdependent effects and that we cannot form a full understanding of their effects by considering them only in isolation. Although the imbalance problem can be considered a symptom of the small disjuncts problem which is solved by using larger training sets, the overlap problem is of a fundamentally different character and the performance of learned classifiers can actually be made worse by using more training data when overlap is present. We also examine the effects of overlap and imbalance on the complexity of the learned model and demonstrate that overlap is a far more serious factor than imbalance in this respect.

1 Introduction

The imbalance problem occurs when the available training data contains significantly more representatives from one class compared to the other. Many classifiers have been shown to give poor performance in identifying the minority class in cases where there is a large imbalance [1]. The machine learning literature acknowledges the imbalance problem as a major obstacle to building accurate classifiers and there has been significant effort invested in searching for solutions [2,3,4,5,6], as well as some investigation into possible root causes of the problem itself [2]. Recent work shows that imbalance is not a problem when the overall size of the training set is sufficiently large [7,8]. These findings suggest that it is instead the problem of small disjuncts—exacerbated by imbalance and small training sets—which is the real cause of poor performance in these cases.

The overlap problem occurs when a region of the data space contains a similar number of training data for each class. This leads to the inference of near equal estimates for the prior probabilities of each class in the overlapping region and makes it difficult or impossible to distinguish between the two classes. There has been comparatively little work done on the overlap problem [5,9,10]; however, recent findings by Garcia et al. [1] have shown that overlap can play an even larger role in determining classifier performance than imbalance. The performance of

A. Farzindar and V. Keselj (Eds.): Canadian AI 2010, LNAI 6085, pp. 220–231, 2010.

Support Vector Machines (SVMs) in this area is of special interest due to the findings by Japkowicz et al. which suggest that SVMs are not sensitive to the imbalance problem in cases where the classes are separable [6].

Previous investigations of the overlap and imbalance problems have taken place largely in isolation. Although some authors have performed experiments in the presence of both factors, the nature of their interaction is still not well understood. Our work demonstrates that these two problems acting in concert cause difficulties that are more severe than one would expect by examining their effects in isolation. This finding demonstrates that we cannot achieve a full understanding of these problems without considering their effects in tandem.

In this paper we provide a systematic study of the interaction between the overlap and imbalance problems as well as their relationship to the size of the training set. We outline a method for testing the hypothesis that these two factors influence classifier performance independently and show through experimentation that this hypothesis is false for the SVM classifier. Finally, we illustrate a connection between model complexity and model performance in the presence of overlap and imbalance.

We have chosen to focus our investigation on SVMs since they have been shown to be particularly robust in the presence of the factors we wish to investigate; however, since our method does not rely on the particulars of the SVM formulation we expect the results reported here to generalize to other classification algorithms. In fact it is likely the case that the interdependence of overlap and imbalance is even stronger in algorithms which are more sensitive to these factors.

2 Detection of Interdependence

In this section we outline a method to test the hypothesis that overlap and imbalance have independent effects on classifier performance. Let us take μ as a measure of overlap between the classes and α as a measure of the between class imbalance[1]. If these two factors act independently we would expect the performance surface with respect to μ and α to follow the relation

$$\mathrm{d}P(\mu, \alpha) = f'(\mu)\,\mathrm{d}\mu + g'(\alpha)\,\mathrm{d}\alpha, \tag{1}$$

where f' and g' are unknown functions. That is, we would expect the total derivative of performance to be separable into the components contributed by each of μ and α. This hypothesis of independence leads us to expect that we can consider the partial derivatives independently, i.e.

$$\frac{\partial}{\partial \mu} P = f'(\mu), \tag{2}$$

$$\frac{\partial}{\partial \alpha} P = g'(\alpha). \tag{3}$$

[1] We provide a concrete method for assigning values to μ and α in Sect. 3, but for the moment we leave the details of this assignment intentionally vague.

The functions f' and g' may not have simple or obvious functional forms meaning that we cannot compute f' and g' directly; however, if f' and g' were known we could find a predicted value for $P(\alpha, \mu)$, up to an additive constant, by evaluating

$$P(\mu, \alpha) = \int f'(\mu)\, \mathrm{d}\mu + \int g'(\alpha)\, \mathrm{d}\alpha + C. \tag{4}$$

Specific values for $P(\mu, \alpha)$ can be computed numerically by training a classifier on a data set with the appropriate level of overlap and imbalance. This requires the use of synthetic data sets since there is no general method to measure the level of class overlap in real data. The use of synthetic data allows us to ensure that other confounding factors, such as problem complexity, remain constant throughout our tests.

Since we expect the partial derivatives of $P(\mu, \alpha)$ to be independent we can compute values for f' by evaluating $P(\mu, \alpha)$ for several values of μ while holding α constant and taking a numerical derivative. Values for g' can be computed in a similar manner by holding μ constant and varying α. These values can then be combined into predicted values for $P(\mu, \alpha)$ using (4). Comparing the predicted values for $P(\mu, \alpha)$ to the observed values will allow us to determine if our hypothesis of independence is sound.

The above method only estimates $P(\mu, \alpha)$ up to the additive constant C. To obtain a value for C we simply need to compute the value of $P(\mu, \alpha)$ for a single point where we have computed both f' and g'.

3 Experiment

In this section we present an experiment designed to test the hypothesis of independence using the procedure outlined in Sect. 2. The data sets we generate for this experiment are a collection of "backbone" models in two dimensions. To generate a data set we sample points from the region $[0, 1] \times [0, 1]$. The range along the first dimension is divided into four regions with alternating class membership (two regions for each class) while the two classes are indistinguishable in the second dimension. To change the overlap level of the classes we allow adjacent regions to overlap. The overlap level is parametrized such that when $\mu = 0$ the two classes are completely separable and when $\mu = 1$ both classes are distributed uniformly over the entire domain. Changing the imbalance level is done by sampling more data points from one class than the other. The imbalance level is parameterized such that α is the proportion of the total data set belonging to the majority class. The total number of samples is kept fixed as α varies. Some example data sets are shown in Fig. 1. These domains are simple enough to be readily visualized yet the optimal decision surface is sufficiently non-linear to cause interesting effects to emerge.

We measure classifier performance over three collections of data sets generated in the manner described above.

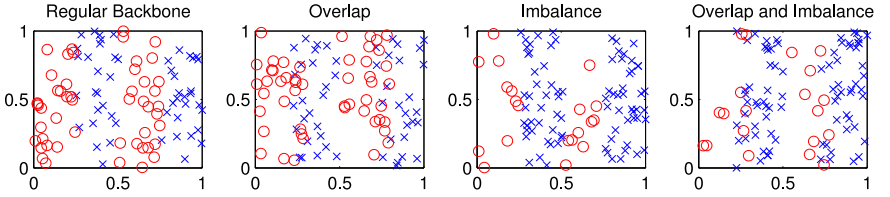

Fig. 1. Sample Data Sets

1. **Varying Overlap** — For these data sets we fix $\mu = 0$ value and α varies over the range $[0.5, 0.95]$.
2. **Varying Imbalance** — For these data sets we vary μ over the range $[0, 1]$ with $\alpha = 0.5$ fixed.
3. **Varying Both** — For these data sets we vary μ and α simultaneously over the ranges $[0, 1]$ and $[0.5, 0.95]$ respectively.

Evaluating our classifier on the first two collections gives us enough information to evaluate (4). Comparing this with the results generated by testing on the third collection of data sets will allow us to determine if overlap and imbalance have independent effects on classifier accuracy.

For each level of imbalance and overlap we measure the classifier performance using several different training set sizes. To build a training set we first select the overlap and imbalance levels as well as the size of the training set and then sample the selected number of points according to the generative distribution defined by the chosen overlap and imbalance. All of our tests are repeated for several training sets sizes varying from 25 to 6400 samples. Testing is done by generating new samples from the same distribution used for training.

We assess classifier performance using the F_1-score of the classifier trained on each data set where the minority class is taken as the positive class. The F_1-score is the harmonic mean of the precision and recall of a classifier and is a commonly used scalar measurement of performance. Our choice of positive class reflects the state of affairs present in many real world problems where it is difficult to obtain samples from the class of interest. The F_1-score is one of the family of F_β-scores which treats precision and recall as equally important.

4 Results

We trained several SVM classifiers on the data sets described in Sect. 3. For each level of overlap, imbalance and training set size we built several classifiers in order to track the variance as well as the overall performance. Parameter values for the SVMs were chosen by selecting a few data sets from our domain of interest and running simulated annealing following the method described in [11] to select the optimal parameters. The optimal parameter values from these tests showed very little variation so we selected a constellation of representative values and left them unchanged for all of our tests. We used the SVM Radial Basis Function kernel for all our tests since it is the most popular non-linear SVM kernel used in practice.

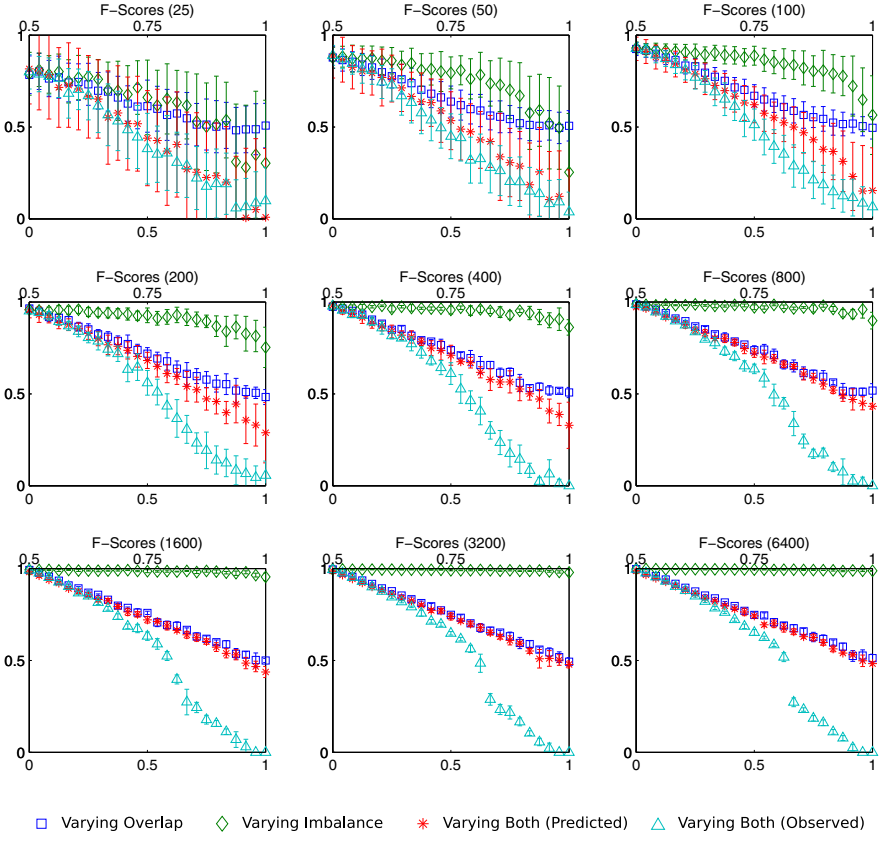

Fig. 2. F_1-scores of several SVM classifiers at different data set sizes. The lower horizontal axis shows the level of overlap and the upper horizontal axis shows the level of imbalance. The vertical axis shows the corresponding F_1-score. Error bars show one standard deviation around the mean.

The performance results from our experiments are shown in Fig. 2. These results clearly show that when the training set size is large the performance predicted by assuming that overlap and imbalance are independent is very different than what is observed. On the other hand, when the training set is small our model is quite accurate, showing only a minor deviation from the observed results.

When the training set size is reasonably large we observe that the class imbalance has very little effect on the classifier performance. This result agrees with previous investigations [6] which suggested that SVMs are not sensitive to class imbalance. When the imbalance level is very high, or when there are few training examples, we still see a drop in performance. This is what we would expect from the existence of small disjuncts in these domains [7,9].

In addition to the F_1-scores we also recorded the number of support vectors from each run. These data are recorded in Fig. 3. Our model of independence

does not make predictions about the number of support vectors so these results cannot be used to test our hypothesis; however, the number of the support vectors can be used as a measure of model complexity.

These results also support the idea that SVMs are not significantly effected by class imbalance when there is sufficient training data. When only imbalance is present we observe that a very small proportion of the total training data is retained as support vectors. This indicates that the SVM has found a highly parsimonious model for the data and, since the corresponding F_1-scores in Fig. 2 are high, we see that these models generalize well. Conversely, when there is class overlap in the training data the number of support vectors rises quickly. This indicates that the SVM has difficulty finding a parsimonious solution despite the fact that there is no increase in complexity of the optimal decision surface. It is interesting to notice that, provided the training set is sufficiently large, the proportion of the training set retained as support vectors shows very little variation across differently sized training sets. A constant proportion of support vectors corresponds to a massive increase in the complexity of the learned model as the training set size is increased.

5 Analysis

5.1 Is Independence Ever a Good Model?

We mentioned previously that for small training set sizes, as well as for small levels of overlap and imbalance, the performance predicted by our model of independence appears to give good predictions for the observed accuracy. Conversely, for high levels of combined overlap and imbalance the predictions given by our model appear to be very poor. In this section we provide a more systematic assessment of the quality of the independence model to determine when, if ever, it might be reasonable to treat these effects independently.

To assess the quality of our model's predictions we preform a two tailed t-test to determine if our predictions differ significantly from the observed results. For these tests we take as our null hypothesis the assumption that overlap and imbalance influence classifier performance independently and compute when it is possible to reject this hypothesis with $>99\%$ confidence. Results from these tests are shown in Fig. 4.

For the smallest training set size we see no strong evidence to reject our hypothesis of independence; however, when there is sufficient training data we see that it is highly unlikely that our hypothesis of independence is correct. It is interesting to note even for very large training sets there is a region of the parameter space where we cannot confidently reject our hypothesis of independence; however, with a large training set this region is quite small and outside it the evidence against independence is quite strong. We also note that the size of this region decreases as the training set size is increased. This is notable since the performance degradation from overlap alone is decreased by using more training data. If the two factors were independent we would expect the combined performance to be very close to the performance in the presence of overlap alone

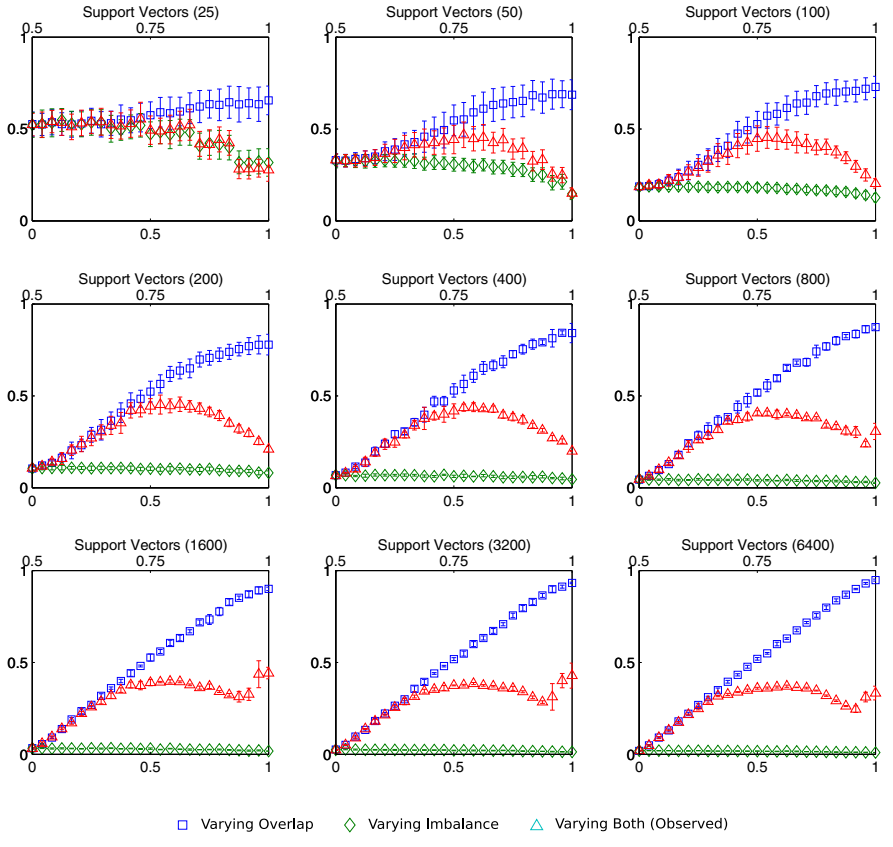

Fig. 3. Proportion of the training set retained as support vectors by several SVM classifiers at different training set sizes. The lower horizontal axis shows the level of overlap and the upper horizontal axis shows the level of imbalance. The vertical axis shows the corresponding proportion of the training set retained as support vectors. Error bars show one standard deviation around the mean.

since with large training sets the degradation from imbalance alone is negligible; however, this is not the case. In light of these observations it is reasonable to conclude that the hypothesis is false in general and we speculate that our inability to reject the model in all cases is merely a case of lack of data.

These results confirm that overlap and imbalance do not have independent effects on performance. We can also see from Fig. 2 that the combined contribution from both factors is made stronger as the training set size increases. Although the contribution from imbalance alone is negligible it cannot be ignored since its presence combined with overlap causes additional degradation beyond the level caused by overlap alone.

From this analysis we see that if independence is ever a good model it can only be when the training set size is very small; however, as we have already

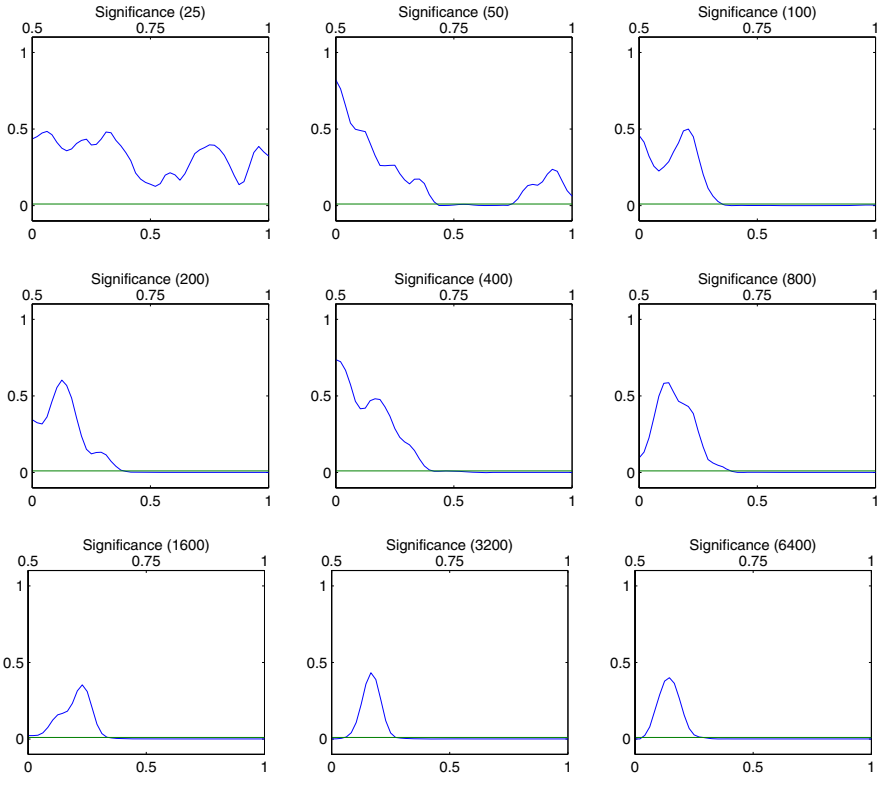

Fig. 4. The p-values for our significance tests. Small p-values indicate a statistically significant deviation between the observed and predicted results for combined overlap and imbalance. The lower horizontal axis shows the level of overlap and the upper horizontal axis shows the level of imbalance. The vertical axis shows the p-value for the associated hypothesis test; also shown is the 99% confidence threshold. These data have been smoothed for readability.

seen, this situation causes other problems and should generally be avoided. We have also not directly shown that the model is good under these conditions, only that we cannot confidently say it is poor.

5.2 Which Is Worse?

We have shown that overlap and imbalance are not independent factors, but the question still remains: Which factor has a more profound effect on the classifier on its own? To answer this we refer again to Fig. 2 which shows classifier performance with overlap and no imbalance as well as with imbalance and no overlap.

When the training set is small, high levels of imbalance cause a dramatic drop in classifier performance; however, with the use of larger training sets this effect

disappears almost entirely. When the training set is large, even an imbalance level of 95% has a barely noticeable effect on performance. This observation is consistent with previous work which showed that problems typically associated with imbalanced data can be better explained by the presence of small disjuncts. As the size of the training set grows the number of training data in each cluster is increased for both the minority and the majority classes. Once there are sufficiently many points in each of the minority class clusters the SVM has no trouble identifying them despite even very high levels of imbalance.

Referring to Fig. 3 we see that more imbalanced training sets actually produce less complex models; i.e. the proportion of the training set retained as support vectors actually drops as the imbalance level is increased. The drop is quite small, however, and the overall proportion of support vectors retained from just imbalance is dwarfed by the proportion retained in the overlapping or combined cases. It is likely that this drop is an artifact of there simply being fewer data available along the margin in the minority class rather than a meaningful reduction in complexity.

Contrasting the above to the effects from overlap, we see from Fig. 2 that overlapping classes cause a consistent drop in performance regardless of the size of the training set. The drop in performance from overlap is linear and performance drops from nearly perfect to $\sim$0.5 as the overlap level is increased. It should be noted that this is exactly what we would expect to happen, even with a perfect classifier. When the classes are overlapping and not imbalanced there are ambiguous regions in the data space where even an optimal classifier with prior knowledge of the generative distributions would not be able to predict the class labels better than chance. The SVM performance in the presence of overlap alone follows exactly the profile we would expect from an optimal classifier in these cases.

It is far more interesting to examine the model complexity in terms of overlap as shown in Fig. 3. Despite the fact that the complexity of the optimal solution remains constant throughout all of our tests, as overlap increases the number of training data retained by the model increases dramatically. This means that although the SVM is able to find a solution which performs comparably to the optimal classifier, the solution it finds becomes progressively more complex as the level of overlap increases.

5.3 Correlating Performance and Model Complexity

We have seen that for sufficiently large training sets there is a sharp drop in performance beyond a certain level of combined overlap and imbalance and that this effect is only seen when both factors are present simultaneously. We also saw that when the two factors are combined the number of support vectors in the resulting model reaches its maximum at an intermediate level of imbalance and overlap. In this section we illustrate the connection between these two observations.

The peak in the number of support vectors (and hence model complexity) is highly correlated with the sharp drop in performance we see with sufficiently

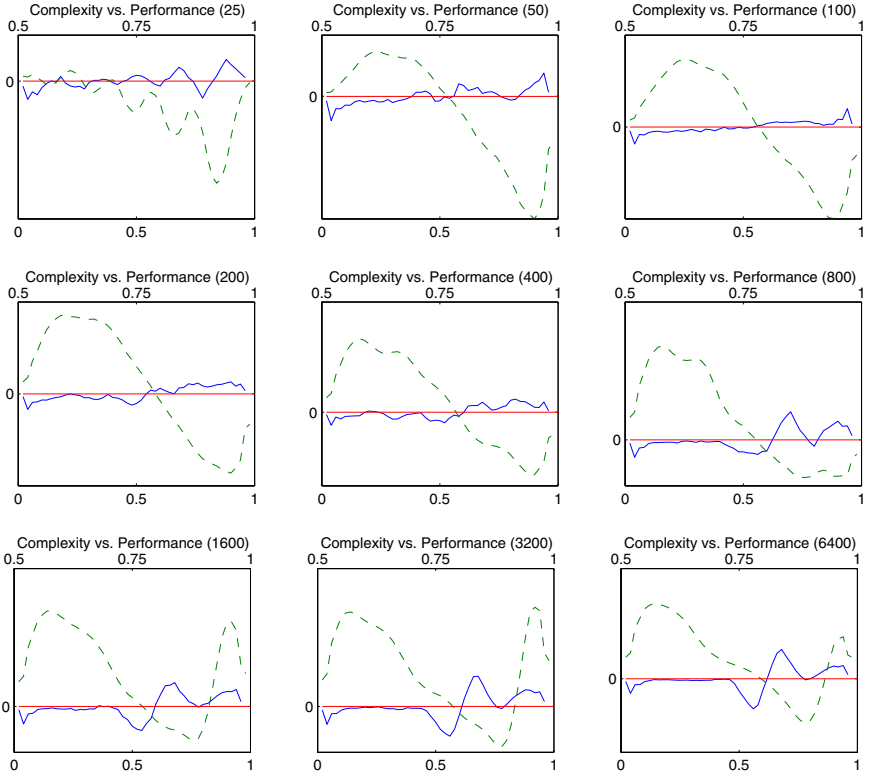

Fig. 5. Comparison between model complexity and performance. The lower horizontal axis shows the level of overlap and the upper horizontal axis shows the level of imbalance. The solid line shows the second derivative of the F_1-score and the dashed line shows the first derivative of the number of support vectors. These data have been smoothed for readability.

large training sets. This correlation is illustrated in Fig. 5 by showing the second derivative of the combined F_1-score and the first derivative of the number of support vectors. The data shown in Fig. 5 has been scaled vertically and smoothed in order to make the plots readable; the important feature to note is where the two lines cross the x-axis. We see that for all but the smallest training set sizes both plots cross the x-axis at approximately $\mu = 0.6$ and $\alpha = 0.78$. The points where the support vector and performance curves cross the x-axis correspond to the peak model complexity and the inflection point in performance respectively. If the overlap and imbalance are increased beyond this point the performance of the trained classifier drops rapidly. Since this effect does not occur when only one of the two factors are present it is clearly an artifact of the combined contribution.

Interestingly, the location of this crossing varies very little as the number of training examples is increased. When there is sufficient training data for the

effect to emerge it is consistently present and its location is relatively unchanged by varying the size of the training set. This suggests that we are observing a type of breaking point—a point where we transition from being able to extract a (somewhat) meaningful representation from the data to a regime where the data representation is not sufficient to build an effective classifier. This observation is supported by a reexamination of Fig. 2 where we can see that performance before the drop is higher in cases where we have used large training sets, but that the performance after the drop is consistently poor regardless of the size of the training set.

6 Conclusion

We have shown that classifier performance varies with overlap and imbalance in a manner that necessitates an interrelationship between these two factors. Comparing the observed performance in cases of combined overlap and imbalance to the performance levels predicted by a model of independence shows that when the two factors are combined the classifier performance is degraded significantly beyond what the model predicts.

Our analysis is consistent with previous results which show that the imbalance problem is properly understood as a problem of small disjuncts in the minority class. When sufficiently many training data are available imbalanced distributions do not impede classification even when the imbalance level is very high. Despite this the imbalance problem cannot be considered solved since levels of imbalance which, in isolation, cause no significant degradation of performance can have a large impact on performance when overlapping classes are also present.

Our analysis of the overlap problem shows that, in isolation, it is a much more serious issue than imbalance. Although the SVM is able to achieve performance comparable to the optimal classifier in the presence of overlap the model complexity tells a different story. Despite the fact that the complexity of the optimal solution remains constant, the complexity of the SVM solution grows proportional to the overlap level and the training set size. This result is important since it shows that more training data—which is often regarded as a panacea for poor performance—can have a detrimental effect on the quality of the learned model.

We have also shown that SVMs have a breaking point where, if the overlap and imbalanced levels are too high we cannot achieve good performance regardless of amount of available training data. We have shown that this breaking point is strongly correlated with the peak model complexity. This effect is notable for several reasons. First, it only appears when both overlap and imbalance are present in tandem, which demonstrates directly that there are effects that we miss by examining overlap and imbalance separately. Second, the insensitivity of this effect to the training set size indicates that it is the result of a systematic weakness of the SVM classifier in the presence of overlap and imbalance rather than a problem with the data. Finally, this finding suggests an avenue for further research into the interaction between overlap and imbalance.

References

1. García, V., Sánchez, J.S., Mollineda, R.A.: An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 397–406. Springer, Heidelberg (2007)
2. Akbani, R., Kwek, S., Japkowicz, N.: Applying support vector machines to imbalanced datasets. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 39–50. Springer, Heidelberg (2004)
3. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations 6(1), 20–29 (2004)
4. Bosch, A.V.D., Weijters, T., Herik, H.J.V.D., Daelemans, W.: When small disjuncts abound, try lazy learning: A case study. In: Seventh Benelern Conference, pp. 109–118 (1997)
5. Yaohua, T., Jinghuai, G.: Improved classification for problem involving overlapping patterns. IEICE Transactions on Information and Systems 90(111), 1787–1795 (2007)
6. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis 6, 429–449 (2002)
7. Japkowicz, N.: Class imbalances: are we focusing on the right issue. In: Workshop on Learning from Imbalanced Data Sets II, pp. 17–23 (2003)
8. Jo, T., Japkowicz, N.: Class imbalances versus small disjuncts. ACM SIGKDD Explorations Newsletter 6(1), 40–49 (2004)
9. Prati, R.C., Batista, G.E.A.P.A., Monard, M.C.: Class imbalances versus class overlapping: An analysis of a learning system behavior. LNCS, pp. 312–321. Springer, Heidelberg (2004)
10. Visa, S., Ralescu, A.: Learning imbalanced and overlapping classes using fuzzy sets. In: ICML 2003 Workshop on Learning from Imbalanced Data Sets II, vol. 3 (2003)
11. Boardman, M., Trappenberg, T.: A Heuristic for Free Parameter Optimization with Support Vector Machines. In: International Joint Conference on Neural Networks, IJCNN 2006, pp. 610–617 (2006)