Using Unsupervised Learning to Guide Resampling in Imbalanced Data Sets

Adam S. Nickerson
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia
Canada B3H 1W5

Nathalie Japkowicz
School of Information
Technology & Engineering
University of Ottawa
Ottawa, Ontario
Canada K1N 6N5

Evangelos Milios
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia
Canada B3H 1W5

Abstract

The class imbalance problem causes a classifier to over-fit the data belonging to the class with the greatest number of training examples. The purpose of this paper is to argue that methods that equalize class membership are not as effective as possible when applied blindly and that improvements can be obtained by adjusting for the within-class imbalance. A *guided resampling* technique is proposed and tested within a simpler letter recognition domain and a more difficult text classification domain. A fast unsupervised clustering technique, Principal Direction Divisive Partitioning (PDDP), is used to determine the internal characteristics of each class. The performance improvement in categories that suffer from a large between-class imbalance (few positive examples) are shown to be improved when using the guided resampling method.

1 INTRODUCTION

The class imbalance problem occurs when there is a large discrepancy between the prior probabilities of the individual classes. That is, one class is represented by a greater number of training examples than the other. [1] If this problem exists within the training data, it can be difficult for a classifier to learn the concept for which there were few examples.

Several methods have previously been proposed to deal with this problem including prior scaling, probabilistic sampling, post scaling [6, Lawrence et al., 1998] and equalizing class membership [5, Kubat and Matwin, 1997]. One shortcoming of these approaches, however, is that they avoid considering the case where, within a single class, the data is distributed according to a mixture density whose components have relative densities that may vary greatly. When faced with such a situation the existing methods that address the class imbalance problem may be counterproductive. While they decrease the difference between the prior probabilities of the classes (the *between-class* imbalance), there is a chance they will increase the difference between the relative densities of the subcomponents within each class (the *within-class* imbalance). Solving one problem by creating another is obviously undesirable.

[1]Throughout this paper, we focus on concept-learning problems in which one class represents the concept at hand (positive class) while the other represents counter-examples of the concept (negative class).

2 THE PROBLEM

As previously observed [8, Mitchell, 1997], the class imbalance problem causes a classifier to over-fit the data belonging to the class with the greatest number of training examples. A simple and effective method for dealing with this problem consists of equalizing class membership by randomly selecting and duplicating examples from the underrepresented class until the two classes are balanced. Although this approach has been shown to increase classification accuracy over that of non-resampling methods [3, Estabrooks, 2000], none of these studies took into consideration the fact that within-class imbalances may occur in addition to between-class imbalances.

The purpose of this paper is to argue that methods that equalize class membership are not as effective as possible when applied blindly and that improvements can be obtained by adjusting for the within-class imbalance.

If we can determine the nature of the subcomponents within each class, we could use that knowledge to guide the resampling. The elements in each subcomponent within each class can then be resampled until each subcomponent has the same number of examples as the largest subcomponent. Then the between-class imbalance can be eliminated by randomly selecting and duplicating members of the underrepresented class (equalizing class membership). This method is hereinafter referred to as *guided resampling.*

We attempt to establish an upper bound of the performance for the guided resampling method by using our prior knowledge of the nature of the subcomponents to guide the resampling as described previously.

In a typical classification problem, we generally would not know the exact partitioning of the subcomponents in advance. In order to guide the resampling, an unsupervised clustering algorithm can be run on each class of the training data in an attempt to find any within-class imbalances. The clusters found are used to guide the resampling as previously described.

3 METHOD

We first employ a method of unsupervised clustering to detect any within-class imbalances in both the positive and negative classes. Using this information, we can avoid increasing the differences in the relative densities of the subcomponents of each class by equalizing the number of members in each subcomponent.

The unsupervised clustering technique, Principal Direction Divisive Partitioning (PDDP), was used to determine the internal characteristics of each class. In our experiments, we used our knowledge of the subcomponents in each class to force PDDP to find that number of clusters within the each class. The clusters were then resampled so that the discovered clusters across both classes each had the same number of examples.

A decision-tree based classifier, C5.0[9, Quinlan, 1998], was trained and used to classify new examples. The results of the guided resampling technique are compared to the results obtained in the absence of a resampling strategy and in the presence of a blind resampling strategy, which resamples at random without taking within-class imbalances into consideration.

3.1 PDDP

The *Principal Direction Divisive Partitioning* (PDDP) [2, Boley, 1997] algorithm operates on a set of m samples where each sample is a vector of n-dimensions containing the attributes of that an example from the training set.

The algorithm determines the internal structure of a class by dividing the set of documents into two clusters by using the principal direction of an $n \times m$ matrix whose i-th column is the vector representing the i-th example. This process is recursively applied to each of the clusters created. The result is a binary tree where the leaf nodes represent the clusters.

PDDP was chosen as the method to determine the internal structure of a class because of it's efficiency. It's expected running time is linear in the number of documents m, modulo the number of iterations with the SVD computation, whereas most clustering algorithms typically have $O(m^2)$ running time.

3.2 Performance Measures

Classification error is not a good performance metric to use when the prior probabilities of the classes differ significantly. [6, Lawrence et al., 1998] When there is a large between-class imbalance, it is trivial to obtain a low error rate simply by classifying all the documents as members of the larger class. Statistics such as *Precision* and *Recall*, two well-known performance metrics within the Information Retrieval community, are not sensitive to this problem.

The *P*recision of a class is the proportion of events labeled as that class which were predicted to be in the class. The *R*ecall of a class is the proportion of correctly detected events which are labeled as that class.

For the purposes of comparison, it is convenient to combine Precision (P) and Recall (R) into a single measure of performance: the *F-measure.* [10, van Rujsbergen, 1979] When Precision and Recall are considered equally important, the F-measure (F) reduces to Figure 1.

$$F = \frac{2PR}{(R+P)}$$

Figure 1: F-Measure

The F-measure lies between zero and one, with values close to one indicating better performance. It is a useful performance metric because it gives low scores to methods that obtain high precision by sacrificing recall or vice versa.

4 EXPERIMENTS

4.1 Letter Classification

To test the practicality of this strategy, we first tested our approach on a simple real-world domain. Using the letter recognition data set available from the UC Irvine Repository, we defined a subtask in which the positive class contained the vowels a and u and the negative class contained the consonants m, s, t and w. Rather than assuming the same number of examples per letter in the training set, we took a subset of the examples for each letter in a way that reflects the letter frequency in English text.[2] While introducing within-class imbalances, this sampling has the advantage of creating a more realistic training set than the one available from the UCI Repository.

In the negative class, the consonants, w is severely underrepresented. If a blind resampling technique is used, there is a good chance that examples of w will not get duplicated often in the resampling process. If we use knowledge of the subcomponent of the negative class, we can ensure that the examples of w get appropriately resampled.

Four experiments were performed on this domain: one with no resampling; one where the between-class imbalance is blindly eliminated; one where PDDP was forced to choose four clusters for the negative class and two clusters for the positive class for the guided resampling process; and one where we use our prior knowledge of the subcomponents of each class of the training set to guide the resampling.

4.1.1 Results

The results from this experiment are reported in Table 1. They indicate that there was no difference in Precision or Recall (and hence no difference in the F-Measure) between the methods of no resampling and blind resampling. When PDDP was used to find the sub-components within each class using the prior knowledge of the actual number of sub-components, slight improvements in Precision and significant improvements in Recall are seen. When we used the prior knowledge of the subcomponents in each class to guide the resampling, it outperforms methods of blind or no resampling but does not perform as well as when the clusters were chosen by PDDP.

Table 1: Results of Letter Classification Experiment

METHOD	P	R	F
No Resampling	0.905	0.818	0.859
Blind Resampling	0.905	0.818	0.859
Guided Resampling (# Clusters Known)	0.923	0.914	0.919
Guided Resampling (Using Known Clusters)	0.935	0.877	0.905

Notice that using either method of guided resampling leads to an improvement in both Precision and Recall.

[2]The following frequencies were used: a: .0856, u: .0249, m: .0249, s .0607, t: .1045, w: .0017. [4, Konheim, 1981] These letters were chosen because their frequencies lead to both between-class and within-class imbalances.

These results served as motivation for trying the guided resampling technique on the more difficult problem of Text Classification.

4.2 Text Classification

The guided resampling technique proposed in the previous section is tested within a text classification domain. More specifically, the problem of classifying an article according to its topic.

The same four experiments were performed on this domain as on the letter classification domain.

4.2.1 Reuters-21578

The Reuters-21578 collection[7, Lewis, 1999] is a collection of 21578 documents originally assembled by Reuters Ltd. in 1987 and later formatted in SGML by David D. Lewis and Stephen Harding. A subset of the Reuters-21578 collection was used to test the aforementioned techniques within the real world domain of text classification.

Specifically, we considered documents that were assigned topics under the categories *earn, acq, money-fx, grain, crude, trade,interest, ship, wheat* and *corn* each of which are represented by a different number of examples as seen in Figure 2.

Table 2: Number of articles for each topic

CATEGORY	NUMBER OF ARTICLES
earn	2709
acq	1488
money-fx	460
grain	394
crude	349
trade	337
interest	289
wheat	198
ship	191
corn	160

The experiment is repeated with each category taking a turn as the positive class. The negative class in each case consists of all the other articles that are not in the positive class.

This text classification domain was initially chosen for our experiments because it was easy to establish an upper bound performance since the sub-components of the negative class are perfectly known. To establish an upper bound on performance for this technique, the prior knowledge of the sub-components of each class is again used to guide the resampling. In an ideal situa-

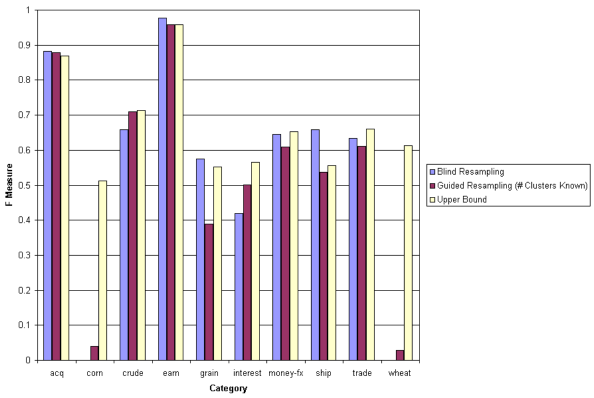

Figure 2: Comparison of F-measure for all categories

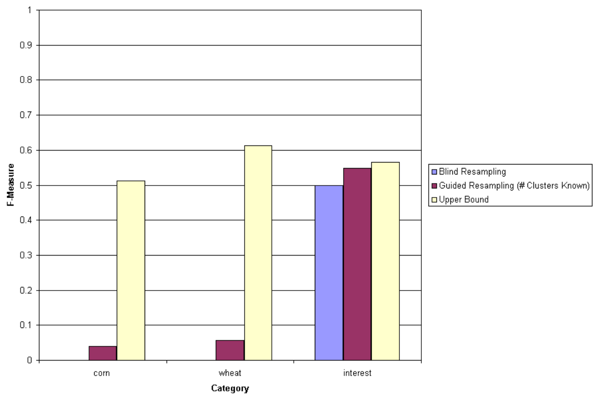

Figure 3: Comparison of F-measures for categories where guided resampling shows improvement

tion, the unsupervised clustering algorithm could perfectly detect these sub-components and that information could be used to guide the resampling accordingly.

Training and testing sets were derived according to the *mod-apte* split [1][Apte, 1994] for the Reuters-21578 collection.

4.2.2 Data Representation

As is standard in text classification experiments, stop words were removed from all documents and the remaining words were stemmed using a Porter stemmer in order to reduce the number of unique words. A feature vector was formed for each document consisting of the counts of the 500 most frequently occurring words (not including the stop words) over the entire document set. This is often referred to as the *bag-of-words* model.

4.2.3 Results

Overall, the results of our method on this domain are not particularly promising. Many of the categories show decreased performance when using guided resampling over blind resampling even when the prior knowledge of the clusters is used. See Figure 2.

Table 3: Average Precision, Recall and F-Measure over all categories

METHOD	P	R	F
No Resampling	0.617	0.394	0.455
Blind Resampling	0.580	0.545	0.560
Guided Resampling (# Clusters Known)	0.650	0.51	0.544
Guided Resampling (# Upper Bound)	0.601	0.751	0.665

Three categories (*corn, wheat* and *Interest*), however, do show improvements in Precision and Recall when using guided resampling (with knowledge of the number of sub-clusters) over blind resampling and this is reflected in their respective F-Measures. See Figure 3.

When using no resampling or blind resampling, no documents were correctly classified for corn or wheat when those categories were acting as the positive class. Using PDDP to find the clusters for guided resampling, C5.0 was able to correctly classify 1 and 2 documents for corn and wheat respectively. When knowledge of the subcomponents is used to guide resampling in these categories, 31 documents are correctly classified as corn and 60 are correctly classified as wheat. When the interest category is considered to be the positive class, guided resampling using PDDP showed improvement over the method of blind resampling.

While using PDDP to guide the resampling does not, in general, achieve results approaching the upper bound, it does achieve better results than when using a blind resampling strategy on these three categories. It is worth noting that these categories suffered some of the greatest between-class imbalances of the entire data set.

5 DISCUSSION

It is likely that the poor results on the Text Classification domain are the result of the representation that we used for the documents. Limiting the feature vectors to the top 500 most frequently occurring words can exclude a lot of relevant information for each document. In that case, documents that share like terms may not be clustered together if those terms are not within the set of words considered for the feature vector.

The improvements seen by using guided resampling on very imbalanced data sets could be applied when methods of blind resampling fail in allowing a classifier to be trained to recognize members of the underrepresented class.

6 FUTURE WORK

Our experiments showed that guided resampling can be useful in the case of severe imbalances. However, to this point, we have assumed that either full knowledge about the subcomponents constituting each class is available or the number of subcomponents in each class is known. The first assumption is very unlikely while the second one is only true in some cases. [3]

An important goal for the future is thus to derive ways to estimate the correct number of subcomponents per class as well as their nature.

It would also be worthwhile to study more rigorously the effects of guided resampling on data where there is very little imbalance. If guided resampling were to be employed when analyzing a new data set whose characteristics are unknown, the random selection of examples from discovered clusters may negatively affect performance.

Once the practicality of our approach is fully established, we would also like to test its generality by applying it with other classification and clustering systems and on other domains where the imbalanced data set problem exists. It would be interesting to determine the effectiveness of this method when using classifiers other than C5.0 such as Multi-Layer Perceptrons based classifiers and when using methods of unsupervised clustering other than PDDP such as k-means clustering or self-organizing maps.

7 CONCLUSION

We have proposed a method for improving methods that deal with the *between-class* imbalance problem by taking any *within-class* imbalances into consideration. These within-class imbalances are detected using *Principal Direction Divisive Partitioning*, an unsupervised clustering algorithm.

The proposed method has shown improvement over existing methods of equalizing class imbalances, especially when there is a large between-class imbalance together with severe imbalance in the relative densities of the subcomponents of each class.

[3] For example, a hospital may know the number of different strains of a bacteria without knowing which patient is affect by which strain.

Acknowledgements

Adam Nickerson would like to acknowledge NSERC for an Undergraduate Student Research Award. Nathalie Japkowicz and Evangelos Milios would like to acknowledge NSERC for their Research Grants.

References

[1] Apte, C., Damerau, F., and Weiss, S. Towards language independent automated learning of text categorization models. In *Proceedings of the 17th Annual ACM/SIGIR conference, 1994.* (1994).

[2] Boley, D. L. Principal direction divisive partitioning. Tech. Rep. TR-97-056, University of Minnesota, Minneapolis, MN, 1997.

[3] Estabrooks, A. A combination scheme for inductive learning from imbalanced data sets. Master's thesis, Dalhousie University, Halifax, Nova Scotia, Canada, 2000.

[4] G. Konheim, A. *Cryptography – A Primer.* John Wiley, 1981.

[5] Kubat, M., and Matwin, S. Addressing the curse of imbalanced training sets: One sided selection. In *14th International Conference on Machine Learning* (San Francisco, CA, 1997), Morgan Kaufmann, pp. 179–186.

[6] Lawrence, S., Burns, I., Back, A., Tsoi, A., and Giles, C. L. Neural network classification and unequal prior class probabilities. In *Tricks of the Trade*, G. Orr, K.-R. Müller, and R. Caruana, Eds., Lecture Notes in Computer Science State-of-the-Art Surveys. Springer Verlag, 1998, pp. 299–314.

[7] Lewis, D. Reuters-21578 text categorization test collection distribution, 1999.

[8] Mitchell, T. M. *Machine Learning.* McGraw-Hill Series in Computer Science. WCB McGraw-Hill, Boston, MA, 1997.

[9] Quinlan, R. Data mining tools see5 and c5.0. Tech. rep., RuleQuest Research, 1998.

[10] van Rijsbergen, C. *Information Retrieval*, 2nd ed. Butterwortths, London, 1979.