Machine Learning from Imbalanced Data Sets 101

Extended Abstract

Foster Provost
New York University
fprovost@stern.nyu.edu

For research to progress most effectively, we first should establish common ground regarding just what is the problem that imbalanced data sets present to machine learning systems. Why and when *should* imbalanced data sets be problematic? When is the problem simply an artifact of easily rectified design choices? I will try to pick the low-hanging fruit and share them with the rest of the workshop participants. Specifically, I would like to discuss what the problem is not. I hope this will lead to a profitable discussion of what the problem indeed is, and how it might be addressed most effectively.

An early stumbling block

A common notion in machine learning causes the most basic problem, and indeed often has stymied both research-oriented and practical attempts to learn from imbalanced data sets. Fortunately the problem is straightforward to fix. The stumbling block is the notion that an inductive learner produces a black box that acts as a categorical (e.g., binary) labeling function.

Of course, many of our learning algorithms in fact do produce such classifiers, which gets us into trouble when faced with imbalanced class distributions. The assumptions built into (most of) these algorithms are:

1. that maximizing accuracy is the goal, and
2. that, in use, the classifier will operate on data drawn from the same distribution as the training data.

The result of these two assumptions is that machine learning on unbalanced data sets produces unsatisfactory classifiers. The reason why should be clear: if 99% of the data are from one class, for most realistic problems a learning algorithm will be hard pressed to do better than the 99% accuracy achievable by the trivial classifier that labels everything with the majority class. Based on the underlying assumptions, *this is the intelligent thing to do*. It is more striking when one of our algorithms, operating under these assumptions, behaves otherwise.

This apparent problem nothwithstanding, it would be premature to conclude that there is a fundamental difficulty with learning from imbalanced data sets. We first must probe deeper and ask whether the algorithms are robust to the weakening of the assumptions that cause the problem. When designing algorithms, some assumptions are fundamental. Changing them would entail redesigning the algorithm completely. Other assumptions are made for convenience, and can be changed with little consequence. So, which is the nature of the assumptions (1 & 2) in question? Investigating this is (tacitly perhaps) one of the main goals of this workshop. However, for many machine learning algorithms it is clear that to a very important extent, the assumptions are made merely for convenience and can be changed easily. Not recognizing this can create misguided research.

In particular, for many machine learning algorithms, the assumptions are embodied to a large extent in the placement of a threshold on a continuous output. Oversimplifying somewhat, let's consider this output to be an estimate of the probability of class membership for a two-class problem. A naïve Bayesian classifer produces such an estimate explicitly. A decision tree does so implicitly, via the class distributions at the leaves. Rule learning programs can make similar estimates. Neural networks produce continuous outputs that can be mapped to probability estimates. If we make the foregoing assumptions (1 & 2), the intelligent choice for classification is to place a threshold of 50% on this output. This typically is what is done by machine learning algorithms; for example, a decision-tree classifier will predict the class most prevalent at the matching leaf.

For learning with imbalanced class distributions, either one or the other of our assumptions (1 & 2) almost certainly will be violated. In some cases there is an important, domain-specific reason why the collection/presentation of examples is skewed toward one class. In other cases, the true class distribution is represented, but the cost of one type of misclassification far outweighs the cost of the other. Neither of these should be a problem. If the true costs and class distributions are known precisely, a simple calculation will produce the correct threshold to use (rather than 50%). If these are not known precisely, how to proceed has been discussed at length elsewhere (e.g., by Tom Fawcett and me (Provost & Fawcett, 1997, 1998, 2000)).

The bottom line is that when studying problems with imbalanced data, using the classifiers produced by standard machine learning algorithms without adjusting the output threshold may well be a critical mistake *(depending on your research question).*

Altering the training balance

A common practice for dealing with imbalanced data sets is to rebalance them artificially. Doing so has been called "up-sampling" (replicating cases from the minority) and "down-sampling" (ignoring cases from the majority). Why should such artificial balancing produce a different classifier than would learning with the original data set? There may be specific reasons

for specific learning algorithms, ***but it is important that the fundamental reasons for different behavior with different stratifications not be confounded with the issue of proper setting of the output threshold (discussed in the previous section)***. I have seen plenty of studies where the "problem" of imbalanced data sets was demonstrated, and then "solved" with up-sampling or down-sampling. Other studies demonstrate that up-sampling or down-sampling does not solve the problem. In most of these studies, it never even was asked whether simply setting the output threshold correctly would be sufficient; without doing so the research may be misleading. (An alternative research methodology is not to consider any particular threshold, but to compare the quality of the models' probability distributions more generally, e.g., across all possible thresholds).

I hasten to add here that I am by no means claiming that there is no benefit to artificially balancing for improving induction. To my knowledge the question is still open as to whether simply changing the distribution skew (without actually looking at *different* data) can improve predictive performance systematically.

On the other hand, there is some evidence that rebalancing the classes artificially in fact does not have a great effect on the predictive performance of learned classifiers. For example, the naïve Bayesian classifier is quite insensitive to stratification. Recent results on decision-tree splitting criteria show that they too are not as sensitive to class skew as conventional wisdom might dictate (Drummond & Holte, 2000).

Designing better sampling strategies

A hope of mine is that from this workshop, and the subsequent research that it stimulates, I will understand more clearly the problem of imbalanced data sets and what to do in their presence. Is the problem simply that there are not enough instances of one class for satisfactory learning? (For extreme imbalances and less-than-massive data sets this certainly is part of the explanation.) Or are there important pathologies of our learning algorithms that become evident when the distributions are highly imbalanced. I will leave the latter to other work, but let me address the former in more detail.

Imbalanced data sets are a problem particularly when there simply are too few data of a certain class, and you end up with problems such as a complete lack of representation of certain important aspects of the minority class. In such cases, it may be useful to look to methods for "profiling" the majority class without reference to instances of the other. I will not address this situation further here either.

There still remains a problem that I believe deserves further investigation, because it has implications for machine learning research more broadly. **What is the appropriate sampling strategy for a given data set?** Is there any reason to believe that the different classes that make up a single learning task will be equally difficult (in terms of the number of instances needed) to learn?

Experience tells us that it is almost always impossible to predict how easy it will be to learn a real-world concept with a complex machine-learning algorithm. However, it may be possible to adjust a sampling strategy dynamically by observing some indicator of the ease of learning the different classes. The most extreme version of this strategy is completely "active" learning: model-based instance selection in an incremental setting, which harkens back to Winston's notion of the "near miss" in the 1970's. I have discussed elsewhere the notion of explicitly considering the search of the example space when designing a learning program (Provost & Buchanan, 1995). What is relevant here is that one can envision a wide variety of policies for doing so, at various points along the spectrum from just taking the distribution as given, all the way to the active selection of each individual instance. If we can understand how machine learning algorithms perform with imbalanced data sets, we can understand how best to operate at certain points along this spectrum. (Note that there is no reason to believe that a 50/50 class distribution would necessarily be best even if the target population is distributed 50/50.)

For example, research may show (I know of no such results yet) that all else being equal--in particular the number of training data--it is better to skew the training distribution toward the class with the larger proportion of small disjuncts. This makes sense, since much research has shown that small disjuncts are more error prone (Weiss, 2000). If this turns out to be so, it may be possible to design a progressive sampling strategy (Provost, Jensen, and Oates, 1999) that decides how next to sample by analyzing the prevalence of small disjuncts in the different classes.

Is understanding these basics really worthwhile?

I have tried to provide rationale here for understanding the basics, including: the need for control conditions for empirical analyses, the desire to identify pathological behaviors of our existing algorithms, the possibility of designing better sampling strategies, as well as the production of new and improved learning algorithms.

A potential criticism of the sampling rationale has some validity and so is worth discussing at this workshop. One might say that typically, in a machine learning setting, *you have the data you have*. Adjusting the class balance by necessity is limited either to replicating the minority class or to throwing away some of the majority class. The former does not add information and the latter actually removes information. Considering this fact, isn't the best research strategy to concentrate on how machine learning algorithms can deal most effectively with whatever data they are given?

There certainly are situations where there is a static, small or moderate-size data set to be used and no opportunity to add to it. However, in my experience, typically either (i) you have far more data than your algorithms can deal with, and you have to select a sample, or (ii) you have no data at all and you have to go through an involved process to create them. In the first case, a simple but vexing practical question is how many data to sample and in what proportion. In the second case, creating data is costly and once again there is the question of how many data to create and in what proportion.

Conclusion

Machine learning from imbalanced data sets is an important problem, both practically and for research. I am confident that developing a clear understanding of this particular problem will have broader-ranging implications for machine learning and AI research. Above I discussed briefly particular interactions with the design of sampling strategies and with the problem of small disjuncts, and I believe that deep understanding of active learning can not be achieved without concomitant understanding of the (simpler) problem of learning with imbalanced data sets. However, I believe also that success at these grander goals will be precarious if we do not first attend to the basics. I thank the workshop organizers for helping to ensure that we do.

Acknowledgements

Thanks to the many with whom I've discussed learning from imbalanced data sets, especially Tom Fawcett.

References

Drummond, C. and R. Holte (2000). "Exploiting the Cost (In)sensitivity of Decision Tree Splitting Criteria." *Proceedings of the Seventeenth International Conference on Machine Learning* (To Appear)

Provost, F. and B. Buchanan, (1995). "Inductive Policy: The Pragmatics of Bias Selection." *Machine Learning* 20, pp. 35-61.

Provost, F. and T. Fawcett (2000). "Robust Classification for Imprecise Environments." To appear in *Machine Learning*.

Provost, F. and T. Fawcett (1998). "Robust Classification Systems for Imprecise Environments." In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*.

Provost, F. and T. Fawcett (1997). "Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions." In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*.

Provost, F., D. Jensen and T. Oates (1999). "Efficient Progressive Sampling." In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*.

G. Weiss (2000). Small Disjuncts: A Research Summary. http://www.cs.rutgers.edu/~gweiss/small_disjuncts.html