Making Evaluation Robust but Robust to What?

Chris Drummond
Institute for Information Technology,
National Research Council Canada,
Ottawa, Ontario, Canada, K1A 0R6
Chris.Drummond@nrc-cnrc.gc.ca

Abstract

Generalization is at the core of evaluation, we estimate the performance of a model on data we have never seen but expect to encounter later on. Our current evaluation procedures assume that the data already seen is a random sample of the domain from which all future data will be drawn. Unfortunately, in practical situations this is rarely the case. Changes in the underlying probabilities will occur and we must evaluate how robust our models to such differences. This paper takes the position that models should be robust in two senses. Firstly, that any small changes in the joint probabilities should not cause large changes in performance. Secondly, that when the dependencies between attributes and the class are constant and only the marginals change, simple adjustments should be sufficient to restore a model's performance. This paper is intended to generate debate on how measures of robustness might become part of our normal evaluation procedures. Certainly some clear demonstrations of robustness would improve our confidence in our models' practical merits.

Introduction

Generalization is at the heart of machine learning research. A core part of evaluation is estimating performance on unseen data. Typically some portion of the data set is held back solely for the purpose of estimating this. Our current evaluation procedures, and the associated theoretical guarantees, assume that the held back data is a random sample of the problem domain. Unfortunately, in practical situations this is rarely the case. Differences in the joint distribution occur over the time from when our algorithm is developed to when it is deployed. Differences occur over the lifetime of the algorithm. Differences occur when the algorithm is deployed in different locations, albeit in the same problem domain. Differences even occur when the domain is static. The people who collected the data we are using for evaluation had various motives none of which are likely to produce random samples. Nevertheless, the data do contain information that is valuable as a source of learning. We can neither afford to throw away the data nor go on assuming that the sample is random. Instead, we must develop new ways of evaluating generalization that account for various degrees of deviation from this assumption.

This paper presents some initial steps of this author's continuing research into why these differences occur, how robustness measures might be included in our evaluation procedures and how classifiers could be made more robust. Many of the differences that will occur depend on the underlying structure of the problem. There is potential here to produce a circularity as the role of learning is to discover this structure. To avoid this, I introduce a few very general problem structures which the users should be able to confidently identify in their problem domains. These structures are much more abstract than would be produced by our learning algorithms. Yet, they considerably constrain the type of changes that are likely to occur between the evaluation stage and when the model is used in practice. The structures presented in this paper by no means exhaust all the possible differences that might be encountered in practice. I would argue, however, it is necessary to begin the process of cataloging the sort of differences that may occur. This will lead to greater confidence in our evaluation procedures and the practical value of our machine learning algorithms.

General Problem Structures

To determine what changes may occur when deploying a model, we need some idea of the causal structure of the problem. It is not the intent, here, to enter the philosophical debate about the nature of causality, but I will make use of two concepts already introduced to the machine learning community by Pearl 1996. Firstly, I use the idea of counterfactuals: what could have happened but didn't. Secondly, I use the idea of causal direction: what proceeds what in causal terms. Counterfactuals indicate what might have been different, here I take this to mean what may be different in the future. The causal direction, the one that defines the relationship between the class and the attributes, will be used to identify the counterfactuals.

To make the point clearer I'll use the medical example of a heart attack (acute myocardial infarction for those more medically inclined). When talking about medical diagnosis, we typically think in terms of the symptoms of a disease. For a heart attack these include chest pain, shortness of breath, nausea, vomiting, palpitations, sweating, and anxiety. In machine learning, heart attack would be the class and the symptoms the attributes. Even if we are not clear about the exact relationship between the attributes and the class,

or among the attributes themselves, there is presumably little doubt on the causal relationship between the disease and the symptoms. The direction is clearly from the disease to the symptoms, the class to the attributes, as shown in figure 1.

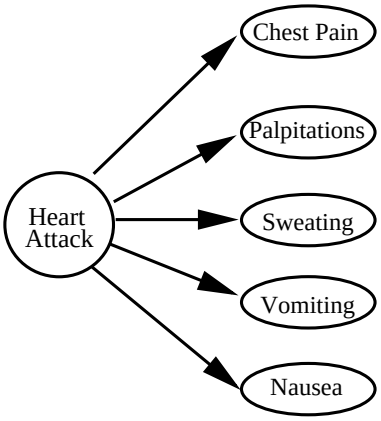

Figure 1: Disease to Symptoms

Suppose we have a patient who is having a heart attack, what are the counterfactuals? Well certainly the heart attack might not have occurred. But given that it did occur, the symptoms are a necessary consequence. This is not a matter of if the symptoms are deterministically or probabilistically related to the disease. What matters is the causal direction. Probabilistically, the marginals associated with the symptoms can't change unless the probability of having a heart attack changes. This is, of course, assuming that dependence structure remains the same. In medical problems, we would expect the relationship between the disease and the symptoms to remain constant over time. What is not guaranteed to remain constant is the frequency of heart attacks in the population. This is one example of a changing class distribution which has been explored by researchers for many different applications. We have ways of evaluating classifiers under these conditions, including ROC (Provost & Fawcett 2001) and cost curves (Drummond & Holte 2006).

We know that there are risk factors for heart attacks including age, smoking and obesity. If we include these factors as attributes, their relationship to the disease will be different from that of the symptoms. The causal direction is from the risk factors to the class to the symptoms, as shown in figure 2.

The risk factors, as well as the symptoms, might form part of the model used to diagnose if someone is having a heart attack or not. A young, trim, non smoker is less likely to be having a heart attack even if he or she is showing some of the symptoms. But for evaluation, we cannot treat these factors as other attributes, they may vary over time. Nowadays, smoking is decreasing yet the population is aging and obesity is rising. How confident could we be in the future efficacy of any predictive model learned using historical data? We should evaluate how robust our models are to changes in these risk factors. We could use bootstrapping (Efron & Tibshirani 1993) controlling the marginals that are expected

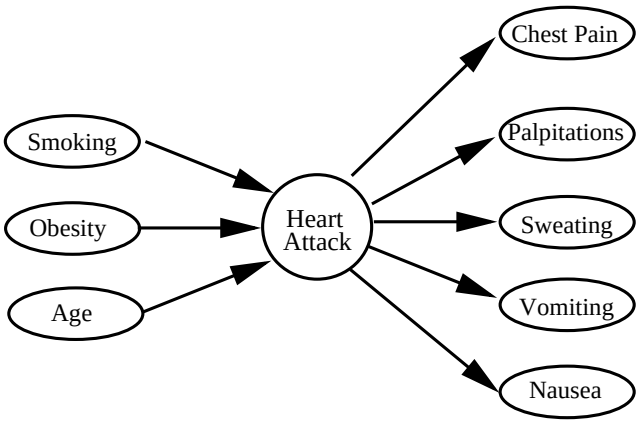

Figure 2: Including Risk Factors

to vary, in the same way that changes in class distribution have been handled (Drummond & Holte 2006). This would give us confidence bounds over any performance measure we might use. We would not only have bounds for sampling variation but also for changes in risk factors. Narrow bounds for both cases would give use high confidence of the robustness of our models and thus their usefulness in the future.

If we are unsure of the causal relationships, we could evaluate our models varying all the attributes. We could bootstrap over all the marginals, assuming there was sufficient variability in the evaluation data set. This may be worth doing for small amounts of variation to be sure that resulting change in performance is also small. But I would argue, it should not be used for testing the robustness to larger changes. One concern is that most existing classifiers would do rather badly. Discriminative classifiers typically have no means of adjustment should all the marginals change. Although simple adjustments may be possible if only some marginals, perhaps just the class distribution (Weiss & Provost 2003), were subject to change.

This would seem to be a strong argument for using generative classifiers, they would surely fare much better. But only if they included the complete dependence structure would they be immune to changes in any marginal. One way for representing this structure, very popular in finance, is the copula (Nelsen 1999). The copula is a cumulative distribution function defined on the unit hypercube. It captures the dependence relationships, perhaps somewhat simplified, between attributes without specifying any of the marginal distributions. So, some models might be robust to changes in all the marginals but most would not. Subjecting them to tests that are more rigorous than needed would be counterproductive.

Another concern is that learning the full dependence structure has many inherent difficulties not the least of which is the small size of many data sets. We also know that in some cases discriminative classifiers have empirically better performance than generative ones (Rubinstein & Hastie 1997; Vapnik 1998). There is also some theory suggesting why this holds true, at least asymptotically (Vapnik 1998). It is quite possible then that learning more than necessary

has a generally downward effect on performance. It would therefore seem reasonable to identify, if at all possible, the attributes causally prior to the class. Then evaluate robustness for major changes only in these attributes.

The aim of the research discussed in this paper is to delineate the ways in which the joint probability distribution could change and yet our classifiers would still be useful. If the changes are too large, we would not expect the model to be of any use. So far I have assumed it is only the marginals of the distribution that might differ. In many domains, such as medical diagnosis, this would seem a reasonable assumption. In other domains, this may not be the case. It still might be possible to define limited, yet commonly occurring, ways the dependence structure might vary and yet our classifiers might still be robust. One idea explored by a number of researchers is concept drift (Widmer & Kubat 1996), which requires that the speed and extent of the drift is limited. Another relevant avenue of research is context sensitivity (Turney 1996), where the importance of attributes or their scaling is situation dependent. With copula, the assumption is that a simple dependence relationship exists between attributes. One that is commonly used, the Gaussian copula, assumes all dependencies are captured by a correlation matrix.We might allow for limited changes in this relationship. For example, we expect changes in the strength of dependence while any independence relationships are maintained.

Whether or not there are other simple constraints on the changes in the joint probability distribution, where robustness might be worth evaluating, remains the subject of future research. I do feel, however, that the effect of small changes in the underlying dependence structure should be evaluated. Again these should not produce large changes in performance.

Conclusion

In this paper, I have argued that models should be robust at least to changes in some marginal probabilities. I believe it is reasonable to have a clear idea of the abstract causal structure in a problem domain; learning would discover the details of this structure. If we can identify those marginals subject to change then we can test the robustness of classifiers to such changes. This would certainly improve our confidence in the long term usefulness of our models.

References

Drummond, C., and Holte, R. C. 2006. Cost curves: An improved method for visualizing classifier performance. *Machine Learning* 65(1):95–130.

Efron, B., and Tibshirani, R. 1993. *An Introduction to the Bootstrap*. London: Chapman and Hall.

Nelsen, R. B. 1999. *An Introduction to Copulas*. Lecture Notes in Statistics. New York: Springer.

Pearl, J. 1996. Causation, action, and counterfactuals. In *Proceedings of the 6th conference on Theoretical Aspects of Rationality and Knowledge*, 51–73. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Provost, F., and Fawcett, T. 2001. Robust classification for imprecise environments. *Machine Learning* 42:203–231.

Rubinstein, Y. D., and Hastie, T. 1997. Discriminative vs informative learning. In *Knowledge Discovery and Data Mining*, 49–53.

Turney, P. 1996. The management of context-sensitive features: A review of strategies. In *Proceedings of the Workshop on Learning in Context-Sensitive Domains,at the 13th International Conference on Machine Learning*, 60–66.

Vapnik, V. 1998. *Statistical Learning Theory*. Wiley.

Weiss, G., and Provost, F. 2003. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research* 19:315–354.

Widmer, G., and Kubat, M. 1996. Learning in the presence of concept drift and hidden contexts. *Machine Learning* 23:69–101.