Machine Learning, 6, 93-98 (1991)

Technical Note
Improved Estimates for the Accuracy of Small Disjuncts

J.R. QUINLAN (QUINLAN@CS.SU.OZ.AU)
Basser Department of Computer Science, University of Sydney, Sydney, Australia 2006

Editor: Tom Dietterich

Abstract. Learning systems often describe a target class as a disjunction of conjunctions of conditions. Recent work has noted that *small* disjuncts, i.e., those supported by few training examples, typically have poor predictive accuracy. One model of this accuracy is provided by the Bayes-Laplace formula based on the number of training examples covered by the disjunct and the number of them belonging to the target class. However, experiments show that small disjuncts associated with target classes of different relative frequencies tend to have different error rates. This note defines the *context* of a disjunct as the set of training examples that fail to satisfy at most one of its conditions. An empirical adaptation of the Bayes-Laplace formula is presented that also makes use of the relative frequency of the target class in this context. Trials are reported comparing the performance of the original formula and the adaptation in six learning tasks.

Keywords. Disjunctive concepts, empirical learning, estimation

1. Introduction

Supervised concept learning from examples plays an important role in current machine learning research. In this area, a system is presented with a substantial number of *training examples*, each from a known *class*. The system then attempts to find definitions of the classes couched in terms used to describe the examples. These definitions should account for the classes of the training examples but, more importantly, they should accurately predict the classes of unseen examples.

A recent paper by Holte, Acker, and Porter (1989) starts with the observation that the class definitions found by learning systems usually consist of several disjuncts. Each disjunct *covers* a subset of the training examples, all or most of which belong to the class associated with the disjunct. The *size* of a disjunct, i.e., the number of training examples that it covers, will generally vary from disjunct to disjunct. Holte et al. focus on *small* disjuncts that cover few training examples and show that the accuracy with which small disjuncts predict the class of unseen examples is much lower than that of their larger brethren. The authors then develop a new bias, *selective specialization*, that improves the predictive accuracy of small disjuncts by making them more specific. The paper contains a careful empirical study demonstrating the benefits of this scheme for small disjuncts.

In deciding whether to further specialize a disjunct on the basis of its size, Holte et al. implicitly assume that all disjuncts of the same size ought to be treated the same way. This note argues that factors other than the size of a disjunct, notably the prevalence of the target class, affect its predictive accuracy.

In many learning tasks, the prevalence of the different classes varies significantly. For example, in one of the learning tasks presented later,[1] 85% of the training examples belong to one class and 15% to the other. Suppose now we have a disjunct that covers only a single training example. The chances are high that our disjunct contains little predictive information as there is scant evidence to support it. If it is used to classify unseen examples drawn from the same population as the training set, and hence with the same class distribution, what accuracy will the disjunct exhibit? If this disjunct is associated with the majority class, it will still be correct much of the time (since most examples that happen to satisfy the disjunct will also belong to the majority class). Conversely, it will be much less accurate if it is identified with the minority class.

For trials with this learning task, Figure 1 shows the average error rate (on unseen examples) of disjuncts that cover up to ten training examples. For smaller sizes, disjuncts of the minority class (denoted by solid circles) are notably less accurate than majority class disjuncts of the same size (denoted by open circles). While disjunct size is clearly important for predicting its accuracy, as found by Holte et al., it is not the end of the story.

In the following sections we present a common estimator of disjunct error and an adaptation that accounts for some of this variation. Empirical studies that compare the original and modified estimator are reported.

2. A simple model

If an event is observed to occur e times in n trials, one estimator for the probability of the event is given by the *Bayes-Laplace* formula

$$\frac{e + 1}{n + 2}$$

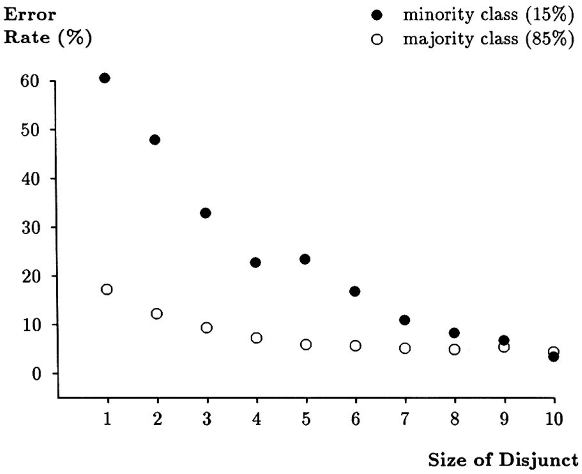

Figure 1. Error rate of small disjuncts associated with different classes.

(Kruskal & Tanur, 1978, p. 256). Analogously, consider a disjunct associated with the target class that is satisfied by n of the training examples, e of which do not belong to the target class. To the extent that we can regard a training example as a trial and an error as an event,[2] the error rate for this disjunct would be given by the same ratio. For a discussion of this use of the Bayes-Laplace formula in estimating error rates, see (Niblett, 1987).

The Bayes-Laplace ratio is derived under the assumption that no information is available regarding the prior probability of error. Suppose, however, that we had reason to believe that the "sample" of n training examples satisfied by this disjunct was drawn from a population containing a proportion C of examples that do not belong to the target class. This proportion, which will be referred to as the *context error* of the disjunct, should clearly influence the predicted accuracy of the disjunct.

We might approximate the context error C by $1 - P(target)$, where $P(target)$ is the prior probability of the target class associated with this disjunct. However, training examples matched by a disjunct are not really sampled from the whole training set; a more relevant collection to regard as a sampling population is the subset of training examples immediately "around" the disjunct, i.e., in the same region of the description space. Since this space is generally non-Euclidean, we take the *context* of a disjunct to be those training examples that fail to satisfy at most one condition of the disjunct. Each training example in this context either matches the disjunct completely or differs from it by a single condition, so the context captures some idea of the locality of the disjunct. If there are n' examples in the context of a disjunct, e' of which do not belong to the class associated with the disjunct, the context error $C = e'/n'$.

Returning to the Bayes-Laplace formula, we see that it can be understood in terms of an *urn model* in which we start with two balls: a white one, representing the target class, and a black one representing other classes. To this are added $n - e$ white balls corresponding to the training examples that belong to the target class and e black balls corresponding to other training examples. The predicted error rate of the disjunct is then the probability of drawing a black ball from this collection.

To adapt this model when the context error is C, it is clear that the initial probability of drawing a black ball should be C. This does not tell us how many balls we should start with—the larger the number, the more slowly the predicted error rate will deviate from this initial probability. Replacing the 2 of the Bayes-Laplace formula with

$$I = \frac{1}{(1 - C)}$$

has proved reasonably satisfactory, although no theoretical justification for this choice can be given. The adapted model of the predicted error rate is thus

$$\frac{e + I \times C}{n + I}$$

Notice that, when the context error C is 0.5, this expression reverts to the Bayes-Laplace formula.

3. Experiments

A series of trials was run with the aim of testing how well this model performs in practice. To simplify matters, learning tasks were restricted to those involving two classes. Six datasets were used:

- *SickEuthyroid* is one aspect of the Garvan thyroid data (Quinlan, Compton, Horn, & Lazarus, 1987). The classes are extremely unbalanced, with 6% in the minority class. Available data was 3772 examples.
- *HighDistinction* concerns A$^+$-level Computer Science students in a database of 373 records. Classes are again very skewed, with 7% in the minority class.
- *LED3* uses a small noisy database from (Breiman, Friedman, Olshen, & Stone, 1984), specialized to the task of distinguishing the digit "3" from other digits. There are 200 examples with 10% in the minority class.
- *ComplexBE* has 11 Boolean attributes A_0, A_1, ..., A_{10}, two of which are irrelevant to the complex concept

  $$(A_1 \vee A_2 \vee A_3) \ \& \ (A_4 \vee A_5) \vee (A_6 \vee A_7) \ \& \ (A_8 \vee A_9).$$

  This concept was included because the exact error rate for any disjunct can be calculated. 1024 examples were used, 15% of them belonging to the minority class.
- *Endgame* involves the chess endgame concept, *King-Knight vs King-Rook lost 3-ply*, for which there are 39 binary-valued attributes and 551 examples. Disjuncts for this class are typically very complicated. The minority class contains 26%.
- *Credit* is a noisy real-world database of 690 records related to the approval of credit applications. There are many irrelevant attributes. The classes are nearly balanced, with 44% in the minority class.

All in all, these learning tasks cover both artificial domains and real-world datasets with a range of difficulties and class imbalances.

In each trial, the available data was divided randomly into a training set of 10% and a test set containing 90% of the examples. (The unusually low proportion of training examples was intended to lead to many small disjuncts.) A decision tree for the training set was generated. Each path from the root to one of the leaves defines a disjunct associated with the class at the leaf, consisting of the conjunction of conditions along the path. The size of this disjunct, its context error over the training set, and its actual error rate on the test set were all determined. For such a disjunct, the *discrepancy* of the model is the absolute difference between the error rate predicted by the model (using only information in the training set) and the actual error rate on the unseen examples in the test set. A similar discrepancy figure was found using the original Bayes-Laplace formula.

Each such trial was repeated 1000 times for every dataset. Discrepancy figures for all disjuncts with size up to 10 were averaged separately for the minority and majority classes. Table 1 shows, for each dataset and both classes, the average discrepancies of the model and the original Bayes-Laplace formula, and the difference between them, with negative differences favoring the model. Due to the large number of repetitions, the standard errors of the averages are small and all differences except the +0.4 are significant.

Table 1. Average discrepancies for the model and Bayes-Laplace formula.

	Minority Class			Majority Class		
Domain	Model	B-L	Δ	Model	B-L	Δ
SickEuthyroid	21.1	42.2	−21.1	24.4	21.4	+3.0
	±0.4	±0.4		±0.3	±0.3	
HighDistinction	18.6	44.5	−25.9	16.0	15.6	+0.4
	±0.4	±0.5		±0.4	±0.3	
LED3	17.2	40.0	−22.8	6.2	16.8	−10.6
	±0.4	±0.7		±0.2	±0.2	
ComplexBE	22.0	29.3	−7.3	11.1	16.5	−5.4
	±0.2	±0.2		±0.1	±0.1	
Endgame	24.8	26.0	−1.2	19.0	21.5	−2.5
	±0.2	±0.2		±0.2	±0.2	
Credit	23.2	29.0	−5.8	29.9	34.1	−4.2
	±0.2	±0.3		±0.3	±0.3	

These figures show that the adapted model is usually more accurate than the original Bayes-Laplace formula. The improvement is marked for domains with more skewed class distributions in which small disjuncts of the minority class tend to have very high error rates; here, the adapted model removes about half of the discrepancy between the actual error rate and that predicted by the Bayes-Laplace formula.

4. Conclusion

Holte et al. (1989) argue with justification that learned concepts must be able to include small disjuncts arising from exceptions and rare cases. Their work offers an approach to reducing the risk of using small disjuncts by changing bias.

This note shows that size alone is not an adequate basis for predicting the error rate of a disjunct, and hence the need for a different bias. In each of the six learning tasks studied, small disjuncts of the majority class proved more accurate than those of the same size associated with the minority class.

The note also presents a more accurate model that employs the idea of the context of a disjunct as the training examples "near" it in the description space. The error rate of training examples in the context is then used as a kind of prior in an adapted form of the Bayes-Laplace formula. The adaptation is ad-hoc rather than theory-based; it would be interesting to see whether a better model could be derived under some set of assumptions about the data and/or the process by which disjuncts are formed. A more thorough model might also take account of factors such as the number of attributes and the number of conditions in the disjunct, both of which are ignored by the present model.

Acknowledgments

Thanks to Teddy Seidenfeld, Rob Holte, and Tom Dietterich for their input. The Machine Learning List, moderated by Mike Pazzani, also provided a useful forum for airing the problem. This research has been supported by a grant from the Australian Research Council.

Notes

1. The *ComplexBE* dataset.
2. This assumption is normally violated because the training set is used in the formulation of the disjunct.

References

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Holte, R.C., Acker, L.E., & Porter, B.W. (1989). Concept learning and the accuracy of small disjuncts. *Proceedings of the llth International Joint Conference on Artificial Intelligence* (pp. 813–818). Detroit, MI: Morgan Kaufmann.

Kruskal, W.H., & Tanur, J.M. (1978). *International encyclopedia of statistics*. New York, NY: Free Press.

Niblett, T. (1987). Constructing decision trees in noisy domains. In I. Bratko and N. Lavrač (Eds.), *Progress in machine learning*. Winslow, U.K.: Sigma Press.

Quinlan, J.R., Compton, P.J., Horn, K.A., & Lazarus, L. (1987). Inductive knowledge acquisition: A case study. In J.R. Quinlan (Ed.), *Applications of expert systems*. Wokingham, U.K.: Addison Wesley.