ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

Coverage-performance estimation for classification with ambiguous data

Thomas Trappenberg *

Dalhousie University - Faculty of Computer Science
6050 University Avenue, Halifax, NS B3H 1W5 - Canada

Abstract. Classifier tradeoffs between accuracy and specificity are often analyzed with *receiver operating curves* (ROC). Here we study a related analysis of the data in terms of *coverage–performance curves* (CPC) which more clearly indicate the presence of ambiguous data in classification problems with overlapping class distributions. We show that feedforward mapping networks are well suited to derive such curves with minimal effort. Based on such classifiers we can identify data that need further analysis before attempting classification with sufficient confidence.

1 Introduction

We often think of classification in situations with well separable data, but overlapping class distributions are common in real world data sets due to noise or missing measurements of features that would uniquely identify the class membership. We refer here to datasets with overlapping class distributions as *ambiguous data* in contrast to *outliers*, which are commonly defined as a subset of observations that appear to be inconsistent with the assumed population of the rest of the observations [1, 2]. There are several problems of concern when applying machine learning classifiers to datasets with ambiguous data including (1) training a classifier may be more difficult than in datasets without ambiguous data, (2) there is an increased risk of overfitting the classifier in order to accommodate ambiguous data, and (3) the overall performance of the classifier, which is commonly reported, does not indicate the range of confidence for classifying each data point.

The performance of classifiers in light of overlapping distributions is classically visualized with a receiver operating characteristic (ROC) curve which specifies the true positive (sensitivity) versus the false positive (1-specificity). ROC analysis is receiving renewed interest specifically in the medical field where 'cautious classification' [3] is mandatory. Provost and Fawcett have shown how ROC analysis can be extended to be robust to imbalanced data [4], and Drummond and Holte have shown that ROC representations can be modivied to include more meaningful cost functions. In this paper we advocate the use of coverage-performance curves (CPC) [5, 6] and the use of classifiers that approximate posteriors, rather than binary decisions, such as probabilistic neural network classifiers (PNNCs) [7, 8, 9].

*I would like to acknowledge valuable discussions with Shunichi Amari and Andrew Back early in this project, and with Saeed Hashemi on applying the methodology to SVMs. This work was supported by NSERC Grant RGPIN 249885-03.

2 Bayesian classification and performance measures with ambiguous data

It is useful to describe classification within a probabilistic framework. Let there be k classes with class labels C_i, each containing data (patterns) that are characterized by a n dimensional feature vector $\mathbf{x}$. The patterns in each class are distributed according to a class probability distribution $p(\mathbf{x}|C_i)$, and data are drawn from each class according to a prior distribution $p(C_i)$. With the knowledge of these distributions we can calculate the posterior distribution $p(C_i|\mathbf{x})$ using Bayes's rule,

$$p(C_i|\mathbf{x}) = \frac{p(C_i)p(\mathbf{x}|C_i)}{\sum_i p(C_i)p(\mathbf{x}|C_i)}, \tag{1}$$

which specifies the probability that a specific pattern $\mathbf{x}$ belongs to the class C_i. The best strategy for predicting the class membership of a given example is to choose the class with the maximal posterior probability for the specific example. The class index of the predicted class is then given by

$$\hat{i} = \arg\max_i p(C_i|\mathbf{x}), \tag{2}$$

where $\arg\max_i$ returns the index of the maximal posterior probability. When classifying data we thus expect a maximal success rate of correct classifications given by

$$p_c = \int_X \max_i p(C_i|\mathbf{x})[\sum_i p(C_i)p(\mathbf{x}|C_i)]d\mathbf{x}, \tag{3}$$

where X is the space of all possible feature patterns. The performance of any classifier is limited by this value. An example for two classes with only one feature value x that consist of two Gaussian distributions with variance $\sigma = 1$, mean $\mu_1 = 1$ for class C_1, and mean $\mu_2 = -1$ for class C_2 is shown in Figure 1A. If we attempt to classify data drawn from these distributions we can achieve a maximal rate of correct classifications following Eqn (3) of

$$p_c = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\mu}{\sqrt{2}\sigma}\right) \approx 0.84. \tag{4}$$

The incorrect classifications are caused by data in the feature region where the different classes have similar posterior probabilities. In many applications, such as clinical classifications, it is mandatory to classify data with a maximal accuracy and confidence level. While the average accuracy might not reach this threshold, it is often the case that many samples can be classified with sufficient confidence. An obvious solution is thus to ignore the difficult regions and to only classify data that can be classified with sufficient confidence. In terms of a Bayesian classification, we only classify data with a posterior probability which exceeds a certain threshold,

$$\hat{i} = \begin{cases} \arg\max_i p(C_i|\mathbf{x}) & \text{if } p(C_{\hat{i}}|\mathbf{x}) > P_t \\ k+1 & \text{elsewhere} \end{cases}, \tag{5}$$

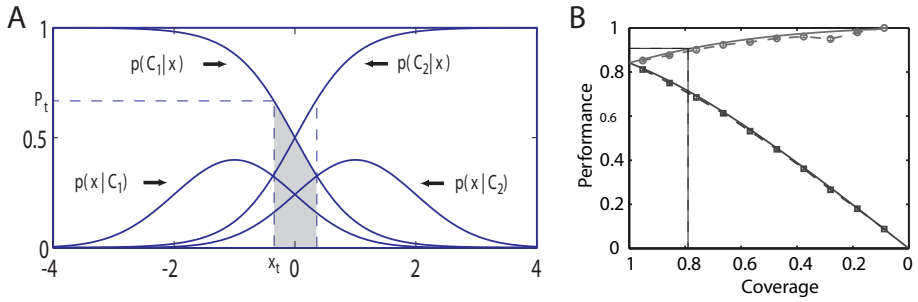

Fig. 1: (A) Example with overlapping Gaussian data. As shown in the text, due to the overlapping class probability density functions, the maximal rate of correct classification in this case is only about 0.84. (B) Coverage–performance curves (CPC) with Gaussian overlapping data for the performance measures P_{high} (upper curve) and P_{low} (lower curve). The solid lines without markers represent the analytical results for these known distributions. The datapoints marked with symbols are results for classifications with probabilistic network.

where P_t is a classification threshold. Examples that can not be classified within this confidence limit are collected in a separate class with index $k+1$, which we call class IDK ('I don't know'). A prediction of a class membership of these examples is left to further analysis, possibly based on further (clinical) tests.

Ignoring some data in the classification prompts the question of how to quantify the performance of the classifier. The answer depends on how much value we place on not classifying some data versus incorrectly classifying examples. We can think of two extreme cases. In one case we can imagine that the classification of all examples is essential, and that not classifying data is to be avoided. A reasonable measure of accuracy is then to take the number of correctly classified examples N_c relative to all the examples in the dataset N and define

$$P_{\text{low}} = \frac{N_c}{N}. \tag{6}$$

In this performance measure, unclassified data are treated as misclassifications. Alternatively, we can choose not to penalize unclassified data by defining P_{high}, which is the fraction of correctly classified examples within the classified examples only (true positive),

$$P_{\text{high}} = \frac{N_c}{N - N_n}, \tag{7}$$

where N_n is the number of examples that were not classified. This definition is suitable for applications in which only the performance of the attempted classifications is important and can be used to determine whether further data collection or testing is required. It is possible to generalize such measures to the confusion matrix in the case of unclassified data and also to more detailed performance measures by including specific, application dependent, cost functions.

A more intuitive way to visualize performance of classifiers with unclassified data compared to ROCs are CPCs. We define the *coverage* as the fraction of classified data relative to the number of all available data in a dataset,

$$c = 1 - \frac{N_n}{N}. \tag{8}$$

The accuracy of the classifier P_{high} is expected to increase when ambiguous data are correctly identified. An increase in the performance value P_{high} with decreasing coverage thus indicate ambiguous data in the datasets. This was demonstrated for classifications with support vector machines in [5]. It is easy to calculate analytically the performance and coverage for different classification thresholds P_t in the above example of the overlapping Gaussian data. These CPCs are shown in Figure 1B as solid lines without markers, the upper curve is for the performance measure P_{high}, the lower curve is for P_{low}. Both curves coincide at $c = 1$. The curve for P_{high} clearly indicates ambiguous data as the performance increases when avoiding classifications of some data. CPCs are useful if an application demands a minimal classification accuracy. For example, if the application demands a minimal accuracy of $P_{\text{high}} = 0.9$, then we know that a little over 20% of the examples can not be classified with sufficient confidence. Our method specifies which data should be eliminated from the classification set.

3 Classification with ambiguous data using probabilistic networks

Machine learning methods to estimate class distributions are needed when the posterior distributions are unknown. Any classifier that can approximate posterior distributions can be used in this approach, but we concentrate here on the application of mapping networks with normalized output values. Such a neural network with probabilistic interpretation can be regarded as a Bayesian classifier. The input layer represents the parameters of the feature vectors where the number of input nodes is equivalent to the number of feature values in the dataset. We use a single hidden layer with sigmoidal activation functions and a output layer with softmax output function

$$\hat{p}(C_i|\mathbf{x}) = \frac{e^{h_i^{\text{out}}}}{\sum_k e^{h_k^{\text{out}}}}, \tag{9}$$

where h_i^{out} is the net input to the output layer. This activation function normalizes the sum of outputs to 1 in order to give them a probabilistic interpretation. The network is trained on the negative cross-entropy

$$E = -\sum_{\mu}\sum_{i} t_i^{\mu} \log\left(y_i(\mathbf{x}^{\mu}, \mathbf{w})\right), \tag{10}$$

which is appropriate in the probabilistic framework [8, 9] using a quasi-Newton optimization algorithm from the NETLAB implementation [10]. The output

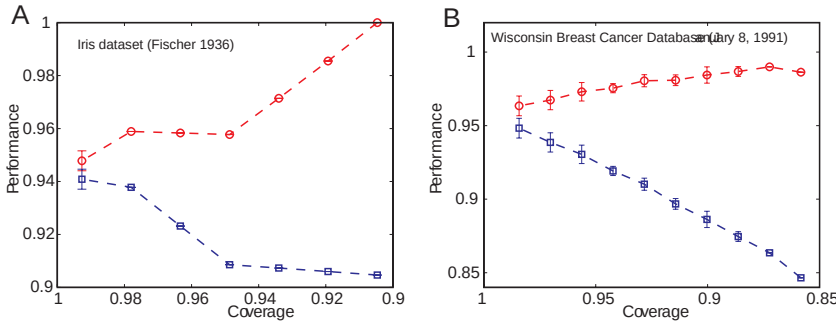

Fig. 2: Coverage–performance curves (CPC) for different datasets derived form probabilistic networks with varying classification thresholds: (A) Wisconsin Breast cancer data and (B) Iris dataset.

value of each individual node is an approximation of the posterior probability for the corresponding class represented by the node for the particular input feature values. It is hence possible to extend the above outlined approach with Bayesian estimates to identify ambiguous data through a variation of the classification threshold. Related approaches have been used in classifications of datasets with outliers [11]. The resulting CPCs in the case of Gaussian example studied in Section 2 are shown in Figure 1B with markers. The neural network classifier was thereby trained on 100 examples and tested on 100 independent examples. The presented results are averages over 100 independent datasets.

To demonstrate the method on real world datasets we applied this method to the Wisconsin Breast cancer data and the Iris dataset from the machine learning repository [12]. The data sets were divided into two equally sized subsets of training data and test data. The resulting CPCs are shown in Figure 2. Both datasets indicate the presence of ambiguous data in these datasets. These curves have been achieved with minimal computational effort and without extensive optimization of the neural network classifier. The Iris dataset shows only a small amount of ambiguous data which agrees with other findings [13]. As expected, when applying the ambiguous data separation algorithm, we achieve 100% classification accuracy on the remaining data. In contrast, the Wisconsin Breast Cancer data show only a small increase in P_{high} with deceasing coverage in agreement with the general finding that this dataset is hard for classification.

4 Conclusions

ROC curves have been used traditionally to compare classifier performances while taking into account tradeoffs between accuracy and specificity. Coverage-performance curves (CPCs) are a more intuitive alternative to quantify how many data have to be omitted in the classification to achieve a specific minimal classification accuracy. Any method that can approximate posterior distribu-

tions form data can be applied to the methods outlined in this paper, but we showed that probabilistic neural networks are well suited for this task. Further studies of CPCs should investigate if they can be used in identify the sources of classification difficulties in different datasets.

Ambiguous data detection and subsequent classification of the remaining data can be done with different classifiers. For example, it is straight forward to use the neural network classifier to identify ambiguous data and then to use a support vector machine (SVM) to do the classification. However, using SVMs for the ambiguous data detection is more problematic. SVMs have been adapted to some extent to novelty detection and ambiguous data identification based on bounded support vectors [14, 5]. However, some preliminary tests on real world data revealed that the estimation of CPCs is extremely difficult with SVMs due to the difficulty in adjusting SVM parameters in order to get appropriate coverage values, and performances on the Gaussian example [5] are considerable purer than the ones reported here.

References

[1] Barnett V. and Lewis T. *Outliers in statistical data*. John Willey & Sons, New York, 3rd edition, 1994.

[2] Ripley B.D. *Pattern recognition and neural networks*. Cambridge Univ. Press, 1996.

[3] C. Ferri and J. Hernandez-Orallo. Coutious classification. *ECAI2004*, 2004.

[4] F. Provost and T. Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. *Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*., 1997.

[5] Hashemi S. and Trappenberg T.P. Using svm for classification in datasets with ambiguous data. *SCI 2002*, 2002.

[6] Thomas P. Trappenberg, Andrew D. Back, and Shun ichi Amar. A performance measure for classification with ambiguous datai. BSIS Technical Report No.99-67, 1999.

[7] Specht D.F. Probabilistic neural networks. *Neural Networks*, 3:109–118, 1990.

[8] D.J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[9] S. Amari. Backpropagation and stochastic gradient desent methods. *Neurocomputing*, 5:185–196, 1993.

[10] I. Nabney. *Netlab: Algorithms for Pattern Recognition*. Springer, 2002.

[11] Sigurdsson S., Larsen J., Hansen L.K., Philipsen P.A., and Wulf H.C. Outlier estimation and detection applicatons to skin leasion classification. *ICASSP'2002, Florida, USA, May 13-17, 2002*, 2002.

[12] C.J. Mertz and P.M. Murphy. UCI repository of machine learning databases.

[13] W. Duch, R. Adamczak, K. Grabczewski, and G. Zal. Hybrid neural-global minimization method of logical rule extraction. *Journal of Advanced Computational Intelligence*, 3:348–356, 1999.

[14] Schoelkopf B., Williamson R., Smola A., Shawe-Taylor J., and Platt J. 2000. *Advances in Neural Information Processing Systems*, 12:582–588, Support Vector Method for Novelty Detection.