EFFECT OF REJECT OPTION ON CLASSIFIER PERFORMANCE

S. Dreiseitl[a], M. Osl[b]

[a]Upper Austria University of Applied Sciences at Hagenberg, Austria
[b]Division of Biomedical Informatics, UCSD, USA

[a]stephan.dreiseitl@fh-hagenberg.at, [b]mosl@ucsd.edu

ABSTRACT

Binary classifier systems that provide class membership probabilities as outputs may be augmented by a reject option to refuse classification for cases that either appear to be outliers, or for which the output probability is around 0.5. We investigated the effect of these two reject options (called "distance reject" and "ambiguity reject", respectively) on the calibration and discriminatory power of logistic regression models. Outliers were found using one-class support vector machines. Discriminatory power was measured by the area under the ROC curve, and calibration by the Hosmer-Lemeshow goodness-of-fit test. Using an artificial data set and a real-world data set for diagnosing myocardial infarction, we found that ambiguity reject increased discriminatory power, while distance reject decreased it. We did not observe any influence of either reject option on the calibration of the logistic regression models.

Keywords: classifier systems, reject option, performance evaluation

1. INTRODUCTION

Decision support systems in biomedicine can augment a physician's diagnostic capabilities by providing an automated second opinion. There are a number of approaches to building such systems, ranging from capturing an expert's domain knowledge in explicit form to using machine learning methods that learn a model from given data without additional human intervention.

Here, we consider only systems of the second kind, and further restrict our attention to models that distinguish between two classes (e.g., classifying cases as either healthy or diseased). Some of these systems, such as logistic regression or neural network models, provide explicit class membership probabilities, i.e., their output is a measure to which degree a case is healthy or diseased. Other machine learning models, such as support vector machines, must be explicitly augmented to provide probability outputs.

The advantages of probability outputs are numerous: Besides facilitating accurate assessments of the system's discriminatory power and calibration via ROC analysis (Bradley 1997; Fawcett 2006) and goodness-of-fit tests (Hosmer et al. 1997; Pigeon and Heyse 1999), probability estimates can also be used for implementing a *reject option*. Such an option allows the system to refrain from making a decision if the predicted membership probability for both classes is around 50%, i.e., if the system cannot make a decision with a reasonable level of certainty.

In addition to rejecting uncertain cases, it may also be desirable for a decision support system to make recommendations only for cases that are similar to the ones that were used for building it. This goal is more difficult to achieve, because it involves estimating how similar a new case is to a set of previously known cases.

In the literature (see Section 2.), the first reject option (around probabilities of 50%) is known as *ambiguity* reject option, and the second (for outlier cases) as *distance* reject option.

In this work, we investigate to which extend the ambiguity and distance reject options have an influence on the quality of a classifier's performance, as measured by its discriminatory power and calibration. For the ambiguity reject option, we use a logistic regression model and do not classify cases for which the model output is close to 0.5. For the distance reject option, we additionally use a one-class SVM to estimate the regions in input space where most of the cases lie. The points outside these regions are then considered to be outliers and rejected from classification.

2. PREVIOUS WORK

The idea of not classifying cases in regions of substantial class overlap, and thus class membership probabilities of around 50%, was proposed by Chow (1970), who was also the first to conduct an investigation into the benefits of the reject rule from a theoretical point of view. Dubuisson and Masson (1993) were the first to consider distance rejection, using nearest neighbor distances to decide whether a point is too far from the remainder of a data set. The particular case of a reject option for nearest neighbor classifiers had been studied earlier by Hellman (1970). Muzzolini et al. (1998) noted that rejection thresholds have to be adjusted to the covariance structure of mixture models to be unbiased, and proposed a method for performing this adjustment. The work of Landgrebe

et al. (2004, 2006) focused on the distance reject option when the classification task is ill-defined in the sense that one clearly defined target class is to be distinguished from another poorly defined class in the presence of an unknown third outlier class. Tax and Duin (2008) proposed a novel method for performing classification with a distance reject option by combining multiple one-class models, one for each of the individual classes.

In the statistical literature, there is some theoretical research on the effects of reject options when rejection costs are different from misclassification costs. The framework of empirical risk minimization provides a theoretical background for the works of Herbei and Wegkamp (2006), Yuan and Wegkamp (2010), and Bartlett and Wegkamp (2008), who derived an SVM classifier with a reject option.

3. METHODS

In this section, we first describe the algorithms we used for building machine-learning models, and then the methods for evaluating the performance of these algorithms.

3.1. Machine learning algorithms

We consider dichotomous classification problems as specified by an n-element data set of m-dimensional input vectors $x_1, \ldots, x_n$ and corresponding class labels $y_1, \ldots, y_n \in \{-1, 1\}$. For logistic regression, we assume there is an additional constant 1 at the first position of the x_i in order to simplify the notation below; these augmented data points are thus $(m+1)$-dimensional.

In a logistic regression model, the optimal values for the $(m+1)$-dimensional parameter vector β are determined by minimizing a negative log-likelihood function. We additionally consider L_2-regularization of logistic regression models by calculating the maximum likelihood estimate β_{ML} as

$$\beta_{\text{ML}} = \arg\min_{\beta} \sum_{i=1}^{n} \log\left(1 + e^{-y_i \beta^T \cdot x_i}\right) + \lambda \beta^T \beta \,.$$

The regularization parameter λ is usually chosen by cross-validation.

The model predictions for new cases x are then given as class-membership probabilities

$$P(y = +1 \,|\, x, \beta_{\text{ML}}) = \frac{1}{1 + e^{-\beta_{\text{ML}}^T \cdot x}} \,.$$

The model outputs are thus logistic transformations of $\beta_{\text{ML}}^T \cdot x$, i.e., values proportional to the distance of x from the hyperplane parameterized by β_{ML}. Ambiguity rejection can therefore be seen to refuse classification for those cases that are within a certain distance from the separating hyperplane.

For distance rejection, we need to estimate the regions of input space in which data are more dense than in others. Standard parametric and non-parametric density estimation algorithms are susceptible to the curse of dimensionality, and therefore not easily applicable in high dimensions. A recent addition to the machine learning arsenal allows us to address this problem without regard to data dimensionality (Schölkopf et al. 2001; Schölkopf and Smola 2002) . *One-class support vector machines* extend standard support vector machine (SVM) methodology to the case of estimating a given fraction $(1-\nu)$ of the support of a data set; the remaining fraction ν are considered outliers.

As with other support vector methods, the one-class SVM algorithm projects the data into a different feature space F using a nonlinear mapping $\Phi : \mathbb{R}^n \to F$. Without a second class, the aim is then to separate the projected data from the origin by as wide a margin ρ as possible. The use of kernel functions k to replace projections and dot product operations is similar to other SVM algorithms. We used a Gaussian kernel

$$k(x_i, x_j) = \exp\Big(-\gamma \left\|x_i - x_j\right\|^2\Big)$$

with inverse variance parameter γ. Values of γ were chosen in such a way that the proportion of outliers was close to ν.

One-class SVMs estimate the data distribution by solving the constrained optimization problem

$$\begin{aligned} \min_{w \in F,\, \xi_i \in \mathbb{R}^m,\, \rho \in \mathbb{R}} \quad & \frac{1}{2}\|w\|^2 + \frac{1}{n\nu}\sum_{i=1}^{n} \xi_i - \rho \\ \text{subject to} \quad & w \cdot \Phi(x_i) \geq \rho - \xi_i, \\ & \xi_i \geq 0 \quad \forall i = 1, \ldots, n \end{aligned}$$

where w is the parametrization of the separating hyperplane in F, and the ξ_i are slack variables. The dual problem is

$$\begin{aligned} \min_{\alpha_i \in \mathbb{R}^n} \quad & \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j \, k(x_i, x_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \frac{1}{n\nu} \quad \forall i = 1, \ldots, n \\ & \sum_{i=1}^{n} \alpha_i = 1 \,. \end{aligned}$$

Support vectors are those data points x_i for which the corresponding α_i satisfies $0 < \alpha_i < \frac{1}{m\nu}$. Outliers are those points for which the decision function

$$f(x) = \sum_{i=1}^{n} \alpha_i \, k(x_i, x) - \rho$$

is negative.

3.2. Evaluation metrics

We used the area under the ROC curve (AUC) as measure of a classifier's discriminatory power, and computed an estimator $\hat{\theta}$ of the AUC via its equivalence to a Mann-Whitney U-statistic as

$$\hat{\theta} = \frac{1}{n_1 \cdot n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \Big(\mathbb{1}\left(p_i^- < p_j^+\right) + \frac{1}{2}\mathbb{1}\left(p_i^- = p_j^+\right)\Big) \,.$$

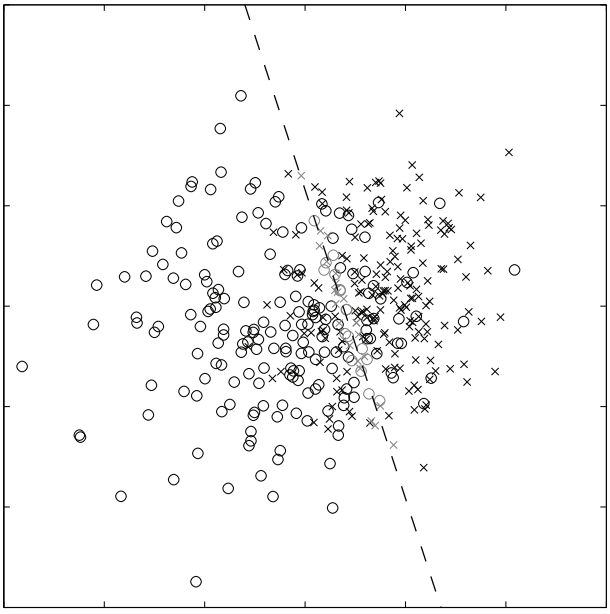

Figure 1: Sample of the artificial data set showing two normally distributed classes with the logistic regression discriminatory line. The points for which no classification is made based on ambiguity rejection are located close to the discriminatory line, and shown in light grey.

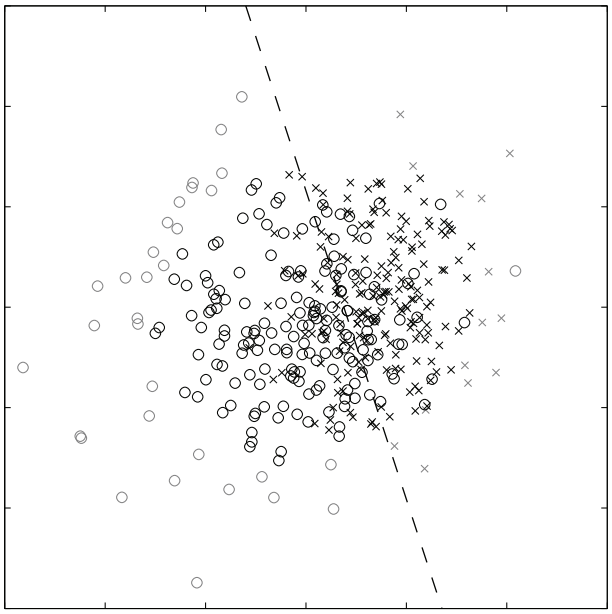

Figure 2: Sample of the artificial data set showing two normally distributed classes with the logistic regression discriminatory line. The points for which no classification is made based on distance rejection are located at the outer edge of the data set, and shown in light grey.

Here, p_i^- and p_j^+ are the classifier outputs for cases from classes -1 and $+1$, respectively, and $\mathbb{1}$ is the Boolean indicator function.

The calibration of a classifier is usually assessed with the Hosmer-Lemeshow C-test (Hosmer and Lemeshow 1980). Although often critizised for a number of drawbacks (Bertolini et al. 2000), it is nevertheless the de-facto standard for determining the goodness-of-fit of a model. As a Pearson chi-squared test, it computes the test statistic

$$C = \sum_{i=1}^{G} \frac{(O_i - E_i)^2}{E_i(1 - \frac{E_i}{n_i})},$$

as the sum of standardized squared differences between the number of observed cases O_i and expected cases E_i for a grouping of classifier outputs into G groups, each with n_i cases. By definition, $G = 10$ and the data is grouped is by sorted classifier outputs. Hosmer and Lemeshow (1980) observed that C has an approximate chi-squared distribution with $G - 2$ degrees of freedom.

4. EXPERIMENTS

Our experiments on the effect of the reject option on classifier performance utilized two data sets, one simple artificial toy problem, and one real-world data set from the domain of predicting acute myocardial infarction.

4.1. Data sets

For the artificial data set, we generated 500 data points each from two multivariate normal distributions with diagonal covariance matrices, each representing one of the two classes. The parameters were chosen to achieve an AUC of about 0.9. A sample of the data with the separating line as determined by logistic regression, along with points in the ambiguity reject region, is shown in Figure 1. The same data, now with points rejected based on distance highlighted, is presented in Figure 2.

The myocardial infarction data set consists of information collected from 1253 patients presenting at the emergency department of the Edinburgh Royal Infirmary in Scotland with symptoms of acute myocardial infarction (AMI). A total of 39 features were recorded, comprising patient data (smoker, diabetes, ...), clinical information (location of pain, sensation of pain, hypertension, ...), and results of an ECG test (LBB, abnormal T wave, ...). To increase diagnostic difficulty, we removed data about ECG measurements, retaining a total of 33 features. The gold standard diagnosis was made by expert physicians based on a combination of blood serum tests with clinical and ECG data. Of the 1253 patients, 274 were diagnosed with AMI, and 979 patients were either declared healthy or to be suffering from other ailments.

4.2. Results

Our experiments were carried out using MATLAB (MathWorks, Natick, MA), with our own implementation of logistic regression models and the libsvm implementation of one-class SVMs (Chang and Lin 2001). For both the artificial as well as the myocardial infarction data set, we trained the logistic regression models using 60% of the data, with the remaining 40% reserved for testing. All data features were normalized to zero mean and unit variance. The experiments were performed 50 times, each

Table 1: Discrimination and classification of logistic regression models on the artificial data set. The results for ambiguity and distance reject options are listed by varying fractions of rejected cases. HL denotes the Hosmer-Lemeshow test statistic; the critical value for $\alpha = 0.05$ is 15.51.

	AUC		HL	
	mean	std	mean	std
logistic regression (baseline)	0.880	0.01	15.07	7.16
ambiguity reject				
$\tau = 0.1$	0.897	0.01	15.22	8.58
$\tau = 0.2$	0.913	0.01	13.93	9.35
$\tau = 0.3$	0.925	0.01	14.73	7.82
$\tau = 0.4$	0.936	0.01	14.82	9.61
distance reject				
$\nu = 0.05$	0.871	0.01	13.93	6.49
$\nu = 0.1$	0.862	0.02	13.10	5.99
$\nu = 0.2$	0.851	0.02	9.65	3.95

Table 2: Discrimination and classification of logistic regression models on the myocardial infarction data set. The results for ambiguity and distance reject options are listed by varying fractions of rejected cases. HL denotes the Hosmer-Lemeshow test statistic; the critical value for $\alpha = 0.05$ is 15.51.

	AUC		HL	
	mean	std	mean	std
logistic regression (baseline)	0.842	0.12	14.51	9.03
ambiguity reject				
$\tau = 0.1$	0.851	0.12	14.10	11.47
$\tau = 0.2$	0.860	0.13	14.29	10.98
$\tau = 0.3$	0.874	0.13	13.14	10.10
$\tau = 0.4$	0.886	0.13	12.19	8.76
distance reject				
$\nu = 0.05$	0.841	0.12	19.38	20.39
$\nu = 0.1$	0.838	0.12	22.61	19.29
$\nu = 0.2$	0.834	0.12	34.81	21.23

with different random allocations of data to the training and test sets. All results are reported as averages and standard deviations on the test set over these 50 runs.

The parameters of our experiments were ν, the fraction of outliers for the distance reject option, and τ, the fraction of cases for the ambiguity reject option. The ambiguity reject cases were the proportion τ of cases for which the model output (class membership probability) was closest to 0.5. A kernel parameter of $\gamma = 0.001$ gave a number of outliers within 10% of the desired value, as specified by ν. The number of support vectors was slightly larger. This is in concordance with theory, which states that ν is an upper bound on the fraction of outliers, but a lower bound on the fraction of support vectors (Schölkopf et al. 2001).

The results of our experiments on the artificial data set are summarized in Table 1. One can observe that ambiguity rejection had a positive effect on AUC. This is to be expected, because ambiguity rejection removes those cases for which most misclassification errors occur. Furthermore, it is also reasonable that the increase in AUC is not as pronounced for larger values of τ, because fewer and fewer ambiguous cases get removed.

On the other hand, there seemed to be no effect of τ on the value of the Hosmer-Lemeshow (HL) test statistic. As is known from the literature (Bertolini et al. 2000), this test dependents strongly on the particular grouping of data points, and showed high volatility in our experiments as well (as indicated by the large standard deviations).

As for distance rejection, the AUC value decreased with increasing numbers of rejected cases, while the HL test statistic showed better model fit. A possible explanation for the first phenomenon is the fact that the rejected points were almost all correctly classified by the model (because they were on the correct side of the discrimination line). Removing them therefore had a detrimental effect on the AUC. The second phenomenon may just be a fluke observation, as evidenced by the high standard deviations and the fact that it was not present in the real-world data.

Table 2 provides the same information for the myocardial infarction data set. Again, we find that ambiguity rejection increased AUC without a clear effect on the HL test statistic (which exhibits even higher standard deviations than on the artificial data set). And again, we observed that distance rejection had a negative effect on AUC. The effect on the HL test statistic was in the opposite direction of the effect it had on the artificial data set.

Distance rejection is also not beneficial if it is performed *before* training, as suggested by Landgrebe et al. (2006). In this case, one rejects data from the training set, and not from the test set. The reasoning for this is that the model may be a better representation of the underlying data generator when outliers are removed prior to model building. Table 3 shows that this is not the case: There was no difference in AUC for the artificial data set, and an even larger negative effect for the real-world data set. There were no discernible effects on the HL test statistics.

5. CONCLUSION

We investigated the effect of the ambiguity and distance reject options on performance of a logistic regression model on an artificial and a real-world data set. We observed ambiguity rejection to increase AUC, and distance rejection to decrease it. Both reject options did not have an effect on classifier calibration.

Table 3: Discrimination and classification of logistic regression models on the artificial and on the myocardial infarction data set, when a fraction ν of cases are removed by distance rejection from the training set prior to model building. HL denotes the Hosmer-Lemeshow test statistic; the critical value for $\alpha = 0.05$ is 15.51.

	AUC		HL	
	mean	std	mean	std
artificial data				
logistic regression (baseline)	0.880	0.01	15.07	7.16
$\nu = 0.1$	0.880	0.01	14.48	6.92
$\nu = 0.2$	0.880	0.01	15.61	7.73
myocard. inf. data				
logistic regression (baseline)	0.842	0.12	14.51	9.03
$\nu = 0.1$	0.828	0.12	15.20	11.75
$\nu = 0.2$	0.818	0.12	14.28	10.40

ACKNOWLEDGEMENTS

This work was funded in part by the Austrian Genome Program (GEN-AU), project Bioinformatics Integration Network (BIN) and the National Library of Medicine (R01LM009520).

REFERENCES

Bartlett, P. and Wegkamp, M. (2008). Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(Aug):1823–1840.

Bertolini, G., D'Amico, R., Nardi, D., Tinazzi, A., and Apolone, G. (2000). One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *Journal of Epidemiology and Biostatistics*, 5(4):251–253.

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159.

Chang, C.-C. and Lin, C.-J. (2001). *LIBSMV: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Chow, C. (1970). On optimum error and reject tradeoff. *IEEE Transactions on Information Theory*, IT-16(1):41–46.

Dubuisson, B. and Masson, M. (1993). A statistical decision rule with incomplete knowledge about classes. *Pattern Recognition*, 26(1):155–165.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Hellman, M. (1970). The nearest neighbor classification rule with a reject option. *IEEE Transactions on Systems Science and Cybernetics*, SSC-6(3):179–185.

Herbei, R. and Wegkamp, M. (2006). Classification with reject option. *Canadian Journal of Statistics*, 4(4):709–721.

Hosmer, D. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10:1043–1069.

Hosmer, D. W., Hosmer, T., le Cessie, S., and Lemeshow, S. (1997). A comparison of goodness–of–fit tests for the logistic regression model. *Statistics in Medicine*, 16:965–980.

Landgrebe, T., Tax, D., Paclík, P., and Duin, R. (2006). The interaction between classification and reject performance for distance-based reject-option classifiers. *Pattern Recognition Letters*, 27(8):908–917.

Landgrebe, T., Tax, D., Paclík, P., Duin, R., and Andrew, C. (2004). A combining strategy for ill-defined problems. In *Proceedings of the 15th Annual Symposium of the Pattern Recognition Association of South Africa*, pages 57–62.

Muzzolini, R., Yang, Y.-H., and Pierson, R. (1998). Classifier design with incomplete knowledge. *Pattern Recognition*, 31(4):345–369.

Pigeon, J. G. and Heyse, J. F. (1999). An improved goodness of fit statistic for probability prediction models. *Biometrical Journal*, 41(1):71–82.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471.

Schölkopf, B. and Smola, A. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.

Tax, D. and Duin, R. (2008). Growing a multi-class classifier with a reject option. *Pattern Recognition Letters*, 29(10):1565–1570.

Yuan, M. and Wegkamp, M. (2010). Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11(Jan):111–130.

AUTHOR BIOGRAPHIES

STEPHAN DREISEITL received his MSc and PhD degrees from the University of Linz, Austria, in 1993 and 1997, respectively. He worked as a visiting researcher at the Decision Systems Group/Harvard Medical School before accepting a post as professor at the Upper Austria University of Applied Sciences in Hagenberg, Austria, in 2000. He is also an adjunct professor at the University of Health Sciences, Medical Informatics and Technology in Hall, Austria. His research interests lie in the development of machine learning models and their application as decision support tools in biomedicine.

MELANIE OSL received her Ph.D. degree from the University of Medical Informatics and Technology (UMIT) in Hall, Austria in 2007. She is currently a postdoc fellow at the Division of Biomedical Informatics at the University of California, San Diego. Her research interests include knowledge discovery and data mining in biomedicine, clinical bioinformatics and machine learning.