Bias-Variance Analysis of Local Classification Methods

Julia Schiffner, Bernd Bischl and Claus Weihs

Abstract In recent years an increasing amount of so called local classification methods has been developed. Local approaches to classification are not new. Well-known examples are the k nearest neighbors method and classification trees (e. g. CART). However, the term 'local' is usually used without further explanation of its particular meaning, we neither know which properties local methods have nor for which types of classification problems they may be beneficial. In order to address these problems we conduct a benchmark study. Based on 26 artificial and real-world data sets selected local and global classification methods are analyzed in terms of the bias-variance decomposition of the misclassification rate. The results support our intuition that local methods exhibit lower bias compared to global counterparts. This reduction comes at the price of an only slightly increased variance such that the error rate in total may be improved.

1 Introduction

Lately the amount of literature on local approaches to classification is increasing. The probably best-known example is the k nearest neighbors method. Many more local approaches have been developed, most of them can be considered localized versions of standard classification techniques such as LDA, logistic regression, naïve Bayes, SVMs, neural networks, boosting etc. The main idea of local classification methods is as follows: Since it is often difficult to find a single classification rule that is suitable for the whole population, rather concentrate on subsets of the population and calculate several individual rules that are only valid for single subsets. Two questions immediately arise: What is the effect of this localization and in which situations are local methods especially appropriate? Many authors when proposing

Julia Schiffner · Bernd Bischl · Claus Weihs
Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany,
e-mail: schiffner@statistik.tu-dortmund.de

a new local classification method just demonstrate superior performance over standard methods on selected data sets. To our knowledge there are only few theoretical results regarding the performance of local methods and no extensive studies that compare several types of local methods across many data sets. A useful concept to gain deeper insight into the behavior of learning algorithms is the bias-variance decomposition of prediction error. It was originally introduced for quadratic loss functions, but generalizations to the misclassification rate have been developed. In a benchmark study on real-world and synthetic data we assess the bias-variance decomposition of the misclassification rate for different types of local methods as well as global methods. For a start in Sect. 2 a short introduction to local approaches to classification is given and different types of local methods are described. The bias-variance decomposition of the misclassification rate and its specific properties are explained in Sect. 3. The benchmark study and the results are presented in Sect. 4. Finally, in Sect. 5 a summary and an outlook to future work are given.

2 Local Approaches to Classification

Due to space limitations we can only give a short introduction. We do not go into details here, but rather point to references on certain topics. In classification it is assumed that each object ω in the population Ω belongs to one and only one class $y = Y(\omega) \in \mathscr{Y}$, with $\mathscr{Y}$ denoting the set of class labels. Additionally, on each object measurements $x = X(\omega) \in \mathscr{X}$ are taken. Both, X and Y, are assumed to be random variables. The aim is to find a classification rule or classifier $D : \mathscr{X} \to \mathscr{Y}$ that based on the measurements predicts the class labels. The set $\mathscr{X}$ is usually called predictor space and often $\mathscr{X} \subset \mathscr{R}^d$. A local classifier is specialized on subsets of the population Ω. A local classification method induces one or more local classifiers and aggregates them if necessary. According to which subsets of the population are addressed several types of local approaches can be distinguished:

Observation-Specific Methods For each object in the population an individual classification rule is built based on the training observations near the trial point x. The best-known example is kNN. But in principle every classification method can be localized this way, for example a localized form of LDA is described in Czogiel et al (2007). While for kNN locally constant functions are fitted to the data, in case of LLDA locally linear functions are used. A review paper concerned with observation-specific methods is Atkeson et al (1997).

Partitioning Methods These methods partition the predictor space $\mathscr{X}$. Examples are CART, mixture-based approaches like mixture discriminant analysis (MDA, Hastie and Tibshirani, 1996) and other multiple prototype methods like learning vector quantization (LVQ, Hastie et al, 2009). Strictly speaking kNN belongs to both groups because it generates a Voronoi tessellation of $\mathscr{X}$.

There are some more local approaches like *multiclass to binary* strategies (Allwein et al, 2000) and *discriminant-adaptive approaches* (Hand and Vinciotti, 2003)

that are beyond the scope of this paper. Since in local classification model assumptions need only be valid for subsets of the population instead of the whole population they are relaxed. For this reason localized methods exhibit more flexibility than their global counterparts and are expected to give good results in case of irregular class boundaries. Localization is only one way to obtain flexible classifiers. Other global ways are e. g. using polynomials of higher degrees and/or kernel methods.

3 Bias-Variance Decomposition of the Misclassification Rate

The bias-variance decomposition of prediction error was originally introduced for quadratic loss. The two main concerns when generalizing it beyond quadratic loss are finding reasonable definitions of bias and variance on the one hand and deriving the decomposition of prediction error on the other hand (e. g. James, 2003).

Definitions of Noise, Bias and Variance Let $\mathscr{P}_{X,Y}$ denote the joint distribution of X and Y and let $P(Y=y|x)$ be the class posterior probabilities. In classification noise at a fixed trial point x is the irreducible or Bayes error

$$Var(Y|x) = E_Y\left[L_{01}\big(Y,S(Y|x)\big)|x\right] = 1-\max_y P(Y=y|x), \text{ where} \tag{1}$$

$$S(Y|x) = \arg\min_y E_Y\left[L_{01}(Y,y)|x\right] = \arg\max_y P(Y=y|x) \tag{2}$$

is the Bayes prediction at x and $L_{01}: \mathscr{Y}\times\mathscr{Y}\to\{0,1\}$ denotes the zero-one loss function. Let $\hat{Y}=D(x)$ denote the prediction at x. Since D is calculated based on training data that result from a random draw from $(\mathscr{X},\mathscr{Y})$ according to $P_{\mathscr{X},\mathscr{Y}}$, $\hat{Y}$ is not fixed. Bias and variance are defined as

$$bias(\hat{Y}|x) = L_{01}\big(S(Y|x),S(\hat{Y}|x)\big) = I\big(S(Y|x)\neq S(\hat{Y}|x)\big), \tag{3}$$

$$Var(\hat{Y}|x) = E_{\hat{Y}}\left[L_{01}\big(\hat{Y},S(\hat{Y}|x)\big)|x\right] = 1-\max_y P(\hat{Y}=y|x), \text{ where} \tag{4}$$

$$S(\hat{Y}|x) = \arg\min_y E_{\hat{Y}}\left[L_{01}(\hat{Y},y)|x\right] = \arg\max_y P(\hat{Y}=y|x) \tag{5}$$

is usually called the main prediction and I is the indicator function. Bias measures the systematic deviation of the main prediction from $S(Y|x)$, while $Var(\hat{Y}|x)$ indicates the random variation of $\hat{Y}$ around the main prediction. S is an operator that reveals the systematic parts of Y and $\hat{Y}$. The definitions given here are natural generalizations of those in the quadratic case, i. e. if the quadratic loss is used instead of L_{01} they reduce to the standard definitions of noise, squared bias and variance.

Decomposition of the Prediction Error In case of quadratic loss the prediction error can be decomposed into the sum of noise, bias and variance and thus both, high bias and high variance, are detrimental to prediction accuracy. In case of zero-one loss the role of variance is completely different. In James (2003) it is shown that in the two-class case the misclassification rate can be decomposed as follows

$$E_{Y,\hat{Y}}\left[L_{01}(Y,\hat{Y})|x\right] = Var(Y|x) + bias(\hat{Y}|x) + Var(\hat{Y}|x) - 2Var(Y|x)bias(\hat{Y}|x)$$
$$-2Var(Y|x)Var(\hat{Y}|x) - 2bias(\hat{Y}|x)Var(\hat{Y}|x) + 4Var(Y|x)bias(\hat{Y}|x)Var(\hat{Y}|x). \quad (6)$$

In contrast to the quadratic case the decomposition additionally contains interactions. The negative interaction effect of bias and variance indicates that variance corrects the prediction in case of bias. If the number of classes is larger than two this correction does not occur for sure, but with a certain probability, which makes the decomposition even more complicated.

Systematic and Variance Effects In order to obtain a simpler decomposition James (2003) distinguishes between bias and variance as measures of systematic deviance and random variation on the one hand and the effects of bias and variance on the prediction error on the other hand. He defines systematic and variance effects as

$$SE(Y,\hat{Y}|x) = E_Y\left[L_{01}\big(Y,S(\hat{Y}|x)\big) - L_{01}\big(Y,S(Y|x)\big)|x\right], \quad (7)$$
$$VE(Y,\hat{Y}|x) = E_{Y,\hat{Y}}\left[L_{01}(Y,\hat{Y}) - L_{01}\big(Y,S(\hat{Y}|x)\big)|x\right]. \quad (8)$$

The systematic effect is the change in prediction error if instead of the Bayes prediction $S(Y|x)$ the main prediction $S(\hat{Y}|x)$ is used. The variance effect measures the change in prediction error due to random variation of $\hat{Y}$ around the main prediction. While $SE(Y,\hat{Y}|x) \geq 0$ the variance effect for the reasons explained above can also take negative values. Under squared loss bias and variance coincide with their respective effects. Generally, an estimator with zero bias also has zero systematic effect and an estimator with zero variance has zero variance effect. With systematic and variance effects an additive decomposition of the misclassification rate is obtained as follows

$$E_{Y,\hat{Y}}\left[L_{01}(Y,\hat{Y})|x\right] = Var(Y|x) + SE(Y,\hat{Y}|x) + VE(Y,\hat{Y}|x). \quad (9)$$

4 Benchmark Study

The aim of our study is to get more insight into the properties of local methods and the effect of localization. As explained in Sect. 2 we assume that local methods in general exhibit rather low bias and decreased bias compared to global counterparts. This reduction probably comes at the price of an increased variance. As explained in Sect. 3 variance needs not be detrimental, but in conjunction with low bias it is likely to be. Questions of interest are: Is the error rate in total increased or decreased and is bias or variance the main contributor to prediction error when using local methods? Moreover, it would be useful to know if there are differences between distinct types of local methods. We expect this since, for example, CART is known for high variance, whereas *k*NN is reported as stable by Breiman (1996). Finally, since we justify our assumption of bias reduction with the increased flexibility of

Table 1 Characteristics of the 26 data sets used in the benchmark study: Number of observations, number of classes, dimensionality, number of numeric and number of categorical predictors

Data Set	#Obs	#Class	Dim	#Num	#Categ	Data Set	#Obs	#Class	Dim	#Num	#Categ
breastcancer	683	2	9	0	9	circle	1000	2	4	4	0
car	1728	4	6	0	6	cuboids	1002	4	3	3	0
crabs	200	2	5	5	0	hvdata	1000	2	6	6	0
credit-g	1000	2	20	7	13	mixture	200	2	2	2	0
crystal	2746	3	37	37	0	orange	1000	2	10	10	0
encoded	794	2	6	6	0	ringnorm	1000	2	4	4	0
glass	214	6	9	9	0	spirals	1000	2	2	2	0
ionosphere	351	2	34	34	0	subclasses	300	2	2	2	0
pima	768	2	8	8	0	subclasses2	990	3	5	5	0
SAheart	462	2	9	8	1	threenorm	1000	2	4	4	0
sonar	208	2	60	60	0	twonorm	1000	2	4	4	0
soybean	562	15	35	0	35	waveform	1000	3	21	21	0
vowel	990	11	9	9	0	xor	1000	8	4	4	0

local methods, we would like to know if there are differences between local methods and global approaches of similar flexibility.

Data Sets We consider 26 data sets, both artificial and real-world data. Most data sets are taken from the UCI repository (Frank and Asuncion, 2010) and the mlbench R-package (Leisch and Dimitriadou, 2010). The threenorm, twonorm and waveform data were used in Breiman (1996). The crystal and encoded data sets are described in more detail in Szepannek et al (2008), the mixture data are taken from Bishop (2006). The hvdata set is an artificial data set described in Hand and Vinciotti (2003). The orange and South-African-heart-disease (SAheart) data are taken from Hastie et al (2009) and the crabs data are available in the MASS R-package (Venables and Ripley, 2002). In Table 1 a survey of the data sets used in the study is given. On the left hand the real-world and on the right hand the artificial data sets are shown. The crabs and wine data sets as well as the hvdata and twonorm data sets pose relatively easy problems since the decision boundary is linear. The circle, orange, ringnorm, subclasses and threenorm data sets are quadratic and the xor and subclasses2 data sets are more complex classification problems.

Classification Methods We consider ten classification methods, three global and seven local methods.

- Global methods: We apply two linear methods, LDA and multinomial regression (MNR), and as a more flexible method a SVM with polynomial kernel (poly SVM). The degree of the polynomial is tuned in the range of 1 to 3.
- Partitioning methods: We consider CART, learning vector quantization (LVQ1) and mixture discriminant analysis (MDA, Hastie and Tibshirani, 1996).
- Observation-specific methods: We use kNN, localized LDA (LLDA, Czogiel et al, 2007) and SVMs with RBF kernel (RBF SVM).

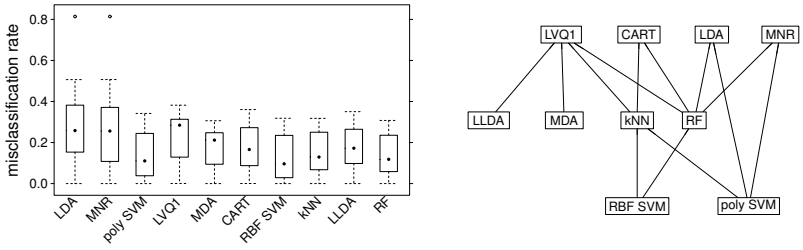

Fig. 1 Left: Boxplot of the error rates on the 26 data sets. Right: Consensus ranking of the classification methods

- Moreover, we consider random forests (RF), which can be regarded as local classification method in conjunction with an ensemble method. Since several classification trees are combined RF normally exhibits lower variance than CART.

Measuring Noise, Bias and Variance In case of artificial data the noise level is usually known. When dealing with real-world data noise can be estimated by means of a consistent classifier. We follow the proposition of James (2003) and employ the 3NN method where the neighbors are weighted by a Gaussian kernel. In order to calculate bias and variance we need to estimate the distribution of $\hat{Y}$. For this purpose we use a nested resampling strategy. In the outer loop we employ subsampling ($4/5$ splits with 100 repetitions). The subsample is used for training while based on the remaining data error, bias, variance and their effects are estimated. In the inner loop we use 5-fold cross-validation for parameter tuning which is required for all employed classification methods except LDA and MNR. A more detailed description of how bias, variance etc. can be estimated can be found in James (2003). All calculations were carried out in R 2.11.1 (R Development Core Team, 2009) using the mlr R-package (Bischl, 2010).

Results While space limitations preclude a full description of the results, some of the main observations are reported here. First, we assess if there are significant performance differences between the classification methods. In Fig. 1 for each classification method a boxplot of the misclassification rates on the 26 data sets is shown. Moreover, a ranking of the methods based on their performance is obtained. For this purpose for each data set a linear mixed effects model is fitted to the misclassification rates obtained on the individual subsamples. Then pairwise differences between the classification methods are tested based on Tukey contrasts and thus an order among the algorithms is established. Finally, a partial consensus ranking of the methods over all data sets is obtained (Eugster et al, 2008). This ranking is not unique. One of eight possible rankings is shown on the right side of Fig. 1. From bottom to top the performance of the classification methods increases. An edge indicates that the error rate of the method displayed on the lower level is significantly smaller than that of the method above. In all eight consensus rankings RBF SVM

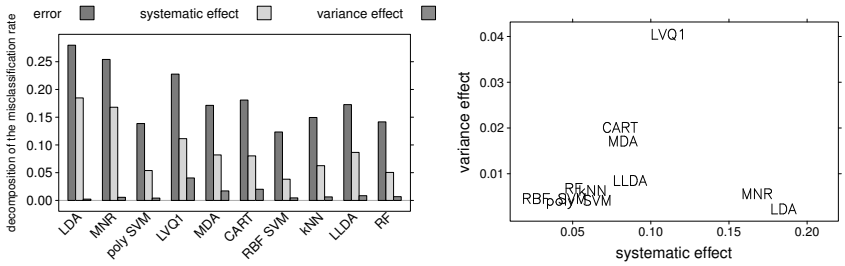

Fig. 2 Left: Error rates, systematic and variance effects averaged over the 26 data sets. Right: Variance effect versus systematic effect

and polynomial SVM show best performance. Moreover, there are no methods that beat LLDA and MDA. Both linear methods and LVQ1 exhibit highest error rates. Next, in order to explain the differences in error rates we consider the bias-variance decomposition. Fig. 2 shows the average misclassification rates as well as systematic and variance effects over all 26 data sets. For all classification methods the systematic effects are considerably larger than the variance effects. Differences in error rates are mainly caused by changes in systematic effect. Both methods with lowest error rates, polynomial and RBF SVMs, exhibit the smallest systematic effects. Their variance effects are hardly larger than that of LDA which is minimal under all classification methods. The localized versions of LDA, LLDA and MDA, both exhibit considerably lower systematic effects, but only slightly increased variance effects. LVQ1 and as expected CART exhibit the largest variance effects. Compared to CART random forests (RF) shows a smaller average variance effect, but as well a smaller systematic effect. In order to visualize which classification methods show a similar behavior with respect to the bias-variance decomposition we plot the variance effect against the systematic effect. We can recognize two clusters, one formed by the linear methods LDA and MNR, the other one formed by highly flexible methods *k*NN, RF as well as RBF and polynomial SVM. Neither global and local methods nor the different types of local methods we mentioned in Sect. 2 can be distinguished based on this plot. The reason may be that the bias-variance decomposition only reflects the degree of flexibility of classification methods and that the way how it is obtained, by localization or otherwise, is not relevant.

5 Summary and Outlook

In order to gain insight into the performance of local classification methods we assessed the bias-variance decomposition of the error rate for local and global classification methods based on real-world and synthetic data. The results support our intuition that localized approaches exhibit considerably lower bias or rather system-

atic effect than global counterparts. Contrary to our assumptions most local methods under consideration, except LVQ1 and CART, only have a slightly increased variance effect, the main contributors to prediction error are noise or systematic effect. In terms of the bias-variance decomposition neither global and local methods nor the different types of local methods could be distinguished. In the future we would like to apply more classification methods. Some types mentioned in Sect.2 like multiclass to binary have not been included yet. Moreover, we plan to relate the results to characteristics of different types of local classification methods. All conclusions were drawn based on averages over the 26 data sets in the study, differences between distinct data sets are not taken into account. In order to gain deeper insight and to get hints in what situation which method will probably perform well it would be useful to relate the results to characteristics of the data sets.

References

Allwein EL, Shapire RE, Singer Y (2000) Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research 1:113–141

Atkeson CG, Moore AW, Schaal S (1997) Locally weighted learning. Artificial Intelligence Review 11(1-5):11–73

Bischl B (2010) mlr: Machine learning in R. URL http://mlr.r-forge.r-project.org

Bishop CM (2006) Pattern Recognition and Machine Learning. Springer, New York

Breiman L (1996) Bias, variance, and arcing classifiers. Tech. Rep. 460, Statistics Department, University of California at Berkeley, Berkeley, CA, URL www.stat.berkeley.edu

Czogiel I, Luebke K, Zentgraf M, Weihs C (2007) Localized linear discriminant analysis. In: Decker R, Lenz HJ (eds) Advances in Data Analysis, Springer, Berlin Heidelberg, Studies in Classification, Data Analysis, and Knowledge Organization, vol 33, pp 133–140

Eugster MJA, Hothorn T, Leisch F (2008) Exploratory and inferential analysis of benchmark experiments. Tech. Rep. 30, Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, URL http://epub.ub.uni-muenchen.de/4134/

Frank A, Asuncion A (2010) UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, URL http://archive.ics.uci.edu/ml

Hand DJ, Vinciotti V (2003) Local versus global models for classification problems: Fitting models where it matters. The American Statistician 57(2):124–131

Hastie T, Tibshirani R (1996) Discriminant analysis by Gaussian mixtures. Journal of the Royal Statistical Society B 58(1):155–176

Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York

James GM (2003) Variance and bias for general loss functions. Machine Learning 51(2):115–135

Leisch F, Dimitriadou E (2010) mlbench: Machine learning benchmark problems. R package version 2.0-0

R Development Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org

Szepannek G, Schiffner J, Wilson J, Weihs C (2008) Local modelling in classification. In: Perner P (ed) Advances in Data Mining. Medical Applications, E-Commerce, Marketing, and Theoretical Aspects, Springer, Berlin Heidelberg, LNCS, vol 5077, pp 153–164

Venables WN, Ripley BD (2002) Modern Applied Statistics with S, 4th edn. Springer, New York, URL http://www.stats.ox.ac.uk/pub/MASS4