Noname manuscript No.
(will be inserted by the editor)

Object selection in credit scoring using covariance matrix of parameters estimations

Alexander A. Aduenko · Anastasia P. Motrenko · Vadim V. Strijov

Received: date / Accepted: date

Abstract We address the problem of outlier detection for more reliable credit scoring. Scoring models are used to estimate the probability of loan default based on the customer's application. To get an unbiased estimation of the model parameters one must select a set of informative objects (customers). We propose an object selection algorithm based on analysis of the covariance matrix for the estimated parameters of the model. To detect outliers we introduce a new quality function called *specificity measure*. For common practical case of ill-conditioned covariance matrix we suggest an empirical approximation of specificity.

We illustrate the algorithm with eight benchmark datasets from the UCI machine learning repository and several artificial datasets. Computational experiments show statistical significance of the classification quality improvement for all considered datasets. The method is compared with four other widely used methods of outlier detection: deviance, Pearson and bayesian residuals and gamma plots. Suggested method performs generally better for both clustered and non-clustered outliers. The method shows acceptable outlier discrimination for datasets that contain up to 30–40% of outliers.

Keywords cash loan · credit scoring · default probability · object selection · outliers filtering.

This publication is funded by the Russian Fond of Basic Research, award number 16-07-01163.

A. Aduenko
Moscow Institute of Physics and Technology
E-mail: aduenko1@gmail.com

Anastasia P. Motrenko
Moscow Institute of Physics and Technology
E-mail: anastasiya.motrenko@phystech.edu

Vadim V. Strijov
Dorodnicyn Computing Centre of RAS
E-mail: strijov@ccas.ru

1 Introduction

We consider the problem of detecting outliers in data samples. Important practical applications of outlier detection can be found in decision problems in the insurance sector, such as designing credit scorecards. The goal of credit scorecards is to estimate the probability of loan default or other potential risks on the basis of customer's application. Design of a scorecard is usually based on analysis of credit histories (Siddiqi 2006). Credit history databases contain millions of entries. Since some entries may be corrupted or contain inadequate values, it is important to select a set of reliable records to construct a high-quality credit scorecard.

Existing approaches to outlier detection split into into direct and indirect methods (Wisnowskia et al. 2001). Direct methods use stepwise add-delete procedures to detect outliers. Indirect methods use special-purpose functions to estimate probability that an object belongs to the sample set. Sebert, Montgomery and Rollier (Sebert et al. 1998) define outliers as non-cluster representatives. To cluster inliers they compute Euclidean distance between objects represented by (predicted value, residual) pairs. Kosinski (Kosinski 1998) presents a comparison of several commonly used techniques for outlier detection, including graphical and analytical methods. The latter include Cook's Square distance (Cook and Weisberg 1989), Mahalanobis Distance
(Rousseeuw and Zomeren 1990) and analysis of residuals (Albert and Chib 1995). Kosinski concludes that outlier detection based on Mahalanobis distance is the most efficient. Hardin and Rocke (Hardin and Rocke 2004) suggest to use the Mahalanobis distance between clusters expected in the sample set. Filzmoser, Maranna and Werner (Filzmoser et al. 2008) develop a method of outlier detection for a sample set in a high-dimensional space.

We use the logistic regression model to estimate the probability of default (Bishop 2006, Bishop and Nasrabadi 2006), since logit function is one of the most common link functions for binary response models (Hahn and Soyer 2005, Hardin and Hilbe 2007). We assume that the feature set (the fields of application form) is fixed and consider sample objects as independent random variables. Application of logistic model to the case of correlated objects is considered, for example, in (Hosmer et al. 2000). To evaluate classification quality we use area under ROC curve (AUC) (Ling et al. 2003). To prove that the classification quality improvement after removing outliers is statistically significant, we sample the AUC distribution and check obtained empirical distribution for normality using ShapiroWilk test (Malkovich and Afifi 1973).

We propose a method of indirect outlier detection. For this purpose, we introduce a specificity measure, which depends on the covariance matrix of the model parameter estimates (Motrenko et al. 2014). We estimate the covariance matrix using approximation of the likelihood function of the regression model around its maximum (Motrenko et al. 2014). The idea of the proposed method comes from the observation that parameter estimates are not robust in the presence of outliers (Croux and Haesbroeck 2003). In this paper we use this property to detect outliers. Leaving out objects from the sample one by one, we estimate regression parameters and analyze the change in parameter estimates. Since outliers have stronger impact on regression parameters then regular objects, deleting an outlier is more likely to result in a considerable change of estimated parameters. This assumption motivated us to introduce a function on objects of the sample called

the specificity measure. Greater change parameter estimates yield higher values of specificity, associated with outliers.

This is not true for sample sets with high percentage of outliers, so the proposed method is valid for sample sets with moderate number of outliers. Moreover, since the method incorporates estimating regression parameters, it is not applicable when the number of features is bigger than the number of objects or when they are close to each other. However, this is not common in credit scoring, since the scoring questionare has 10-100 fields whereas credit stories databases contain thousands or millions of records.

We use data from the UCI machine learning repository to compare the proposed method with alternatives. To provide more extensive comparison, in addition to consumer loans datasets (German dataset 1994, Australian dataset 1987) we use the following benchmark datasets: heart disease in South Africa (SAHD dataset 1993), wine quality (Wine data 1991), yeast dataset (Yeast dataset 1996), breast cancer dataset (Breast cancer dataset 1992), contraceptive usage dataset (Contraceptive dataset 1987) and housing dataset (Housing dataset 1978).

The paper is structured as follows. In Section 2 we formulate the problem of parameter estimation in logistic regression and discuss the properties of the optimization problem that allow us to introduce specificity as measure for outlier detection. In Section 3 we propose a new method of outlier detection and a more practical modification for the proposed method. In Section 4 we provide empirical analysis of the proposed method and compare it to deviance, Pearson and bayesian residuals (Albert and Chib 1995), gamma plots (Evans and Jones 2002).

2 Problem statement

Consider a sample set $D = \{(\mathbf{x}_i, y_i)\}$. Let $\mathcal{I} = \{1, \ldots, m\} \ni i$ denote a set of indices for the sample set. By $\mathbf{X}$ denote the matrix $\mathbf{X} = [\mathbf{x}_1^\intercal, \ldots, \mathbf{x}_m^\intercal]^\intercal$, where m is the size of the set D, and n is the number of features. By $\mathbf{y} = [y_1, \ldots, y_m]^\intercal$ denote a target vector, where $y_i \in \{0, 1\}$; $y_i = 1$ if the i-th record corresponds to the lawn default case and $y_i = 0$ if the i-th record corresponds to the non-default case.

Suppose that y_i is a realization of the Bernoulli random variable $Y_i \sim Be(p_i)$, where p_i is the probability of default. Let the random variables $\{Y_i\}_{i=1}^m$ be mutually independent. Furthermore, assume that p_i is given by logistic regression model:

$$p_i = f(\mathbf{x}_i, \mathbf{w}) = f_i = \frac{1}{1 + \exp\left(-\mathbf{x}_i^\intercal \mathbf{w}\right)}, \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^n$ is a vector of the model parameters. Model parameters $\hat{\mathbf{w}}$ maximize data likelihood

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w} \in \mathbb{R}^n} L(\mathbf{y}|\mathbf{X},\ \mathbf{w}), \tag{2}$$

where $L(\mathbf{y}|\mathbf{X},\ \mathbf{w})$ is the data likelihood function. The problem is to select a subset $(\mathbf{X}_\mathcal{B},\ \mathbf{y}_\mathcal{B})$ of reliable objects

$$\mathbf{X}_\mathcal{B} = [\mathbf{x}_i^\intercal]^\intercal,\ i \in \mathcal{B}, \quad \mathbf{y}_\mathcal{B} = [y_i]^\intercal,\ i \in \mathcal{B},$$

where $\mathcal{B} \subseteq \mathcal{I}$ is the index set of selected objects. An object $\mathbf{x}_i$ is considered an outlier if $i \in \mathcal{I} \setminus \mathcal{B}$. Further we use a cross-validation procedure to estimate the

parameters and select the set $\mathcal{B}$ of indices of non-outliers. Let a random partition $\mathcal{I} = \mathcal{S} \sqcup \mathcal{T}$ of the index set $\mathcal{I}$ split the dataset into training $(\mathbf{X}_\mathcal{S},\, \mathbf{y}_\mathcal{S})$ and testing $(\mathbf{X}_\mathcal{T},\, \mathbf{y}_\mathcal{T})$ samples. The training sample is used for parameters estimation, and the testing sample is used for object selection. The object selection problem is to detect outliers and remove them, maximizing likelihood of the regular data:

$$\mathcal{B} = \underset{\mathcal{B}\subseteq\{1,\ldots,m\}}{\arg\max}\; L\left(\mathbf{y}_{\mathcal{T}\bigcap\mathcal{B}}|\mathbf{X}_{\mathcal{T}\bigcap\mathcal{B}},\, \hat{\mathbf{w}}\right), \tag{3}$$

where

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^n,\, \mathcal{B}\subseteq\{1,\ldots,m\}}{\arg\max}\; L(\mathbf{y}_{\mathcal{B}\bigcap\mathcal{S}}|\mathbf{X}_{\mathcal{B}\bigcap\mathcal{S}},\, \mathbf{w}). \tag{4}$$

Since random variables $\{Y_i\}$, $i \in \mathcal{B}$ are mutually independent, we obtain the following expression for data likelihood

$$L(\mathbf{y}_\mathcal{B}|\mathbf{X}_\mathcal{B},\, \mathbf{w}) = \prod_{i\in\mathcal{B}} f_i^{y_i}\left(1-f_i\right)^{1-y_i},$$

or, equivalently, the following expression for negative logarithm of likelihood function L:

$$l(\mathbf{w}) = -\ln L(\mathbf{y}_\mathcal{B}|\mathbf{X}_\mathcal{B},\, \mathbf{w}) = -\sum_{i\in\mathcal{B}} y_i \ln f_i + (1-y_i)\ln(1-f_i). \tag{5}$$

The second derivative of $l(\mathbf{w})$, or the *hessian* matrix, is given by

$$\mathbf{H} = \sum_{i\in\mathcal{B}} \mathbf{x}_i f_i(1-f_i)\mathbf{x}_i^\intercal = \mathbf{X}_\mathcal{B}^\intercal \mathbf{R}\mathbf{X}_\mathcal{B}$$

Under assumption that the columns of the matrix $\mathbf{X}$ are linearly independent the matrix $\mathbf{H}$ is positive definite and the negative loglikelihood function $l(\mathbf{w})$ is convex (Boyd and Vandenberghe 2004). Therefore, the likelihood function $L(\mathbf{w})$ has unique maximum and the problem (4) is well defined. Note that $\mathbf{H}$ is degenerate if there exists $\mathbf{u} \neq \mathbf{0}$ such that $\mathbf{X}\mathbf{u} = \mathbf{0}$.

Estimation of the model parameters. Since there is no analytical solution for the optimization problem (4) we use the following iterative procedure to estimate $\hat{\mathbf{w}}$.

Let $\hat{\mathbf{w}}_0$ be an initial estimation of parameter vector. Suppose that the $(k-1)$-th iterative approximation $\hat{\mathbf{w}}_{k-1}$ has already been computed. Using Newton-Raphson method (Bishop 2006), we obtain k-th iterative approximation $\hat{\mathbf{w}}_k$ of $\hat{\mathbf{w}}$:

$$\hat{\mathbf{w}}_k = \hat{\mathbf{w}}_{k-1} - \mathbf{H}^{-1}\nabla l(\hat{\mathbf{w}}_{k-1}) = \hat{\mathbf{w}}_{k-1} - \left(\mathbf{X}^\intercal\mathbf{R}\mathbf{X}\right)^{-1}\mathbf{X}^\intercal(\mathbf{f}-\mathbf{y}). \tag{6}$$

Given the initial parameter vector $\hat{\mathbf{w}}_0$ the procedure is well defined. Iterative computation (6) terminates when $\|\hat{\mathbf{w}}_k - \hat{\mathbf{w}}_{k-1}\|_2 < \varepsilon$ for some $\varepsilon > 0$.

Estimation of covariance matrix for $\hat{\mathbf{w}}$. The proposed method for outlier detection involves estimation of the covariance matrix $\mathbf{\Sigma}$ of posterior distribution

$$p(\mathbf{w}|\mathbf{X},\ \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})} \tag{7}$$

of estimated parameters $\hat{\mathbf{w}}$. Suppose the prior distribution $p(\mathbf{w})$ is pseudo uniform. Then, since $p(\mathbf{y}|\mathbf{X})$ is constant with respect to $\mathbf{w}$, posterior distribution $p(\mathbf{w}|\mathbf{X},\ \mathbf{y})$ is proportional to $p(\mathbf{y}|\mathbf{X},\ \mathbf{w})$

$$p(\mathbf{w}|\mathbf{X},\ \mathbf{y}) \propto p(\mathbf{y}|\mathbf{X},\mathbf{w}).$$

To estimate covariance matrix $\mathbf{\Sigma}$ of $p(\mathbf{w}|\mathbf{X},\ \mathbf{y})$ we use Taylor approximation of $l(\mathbf{w}) = \ln p(\mathbf{y}|\mathbf{X},\mathbf{w})$ around its minimum $\mathbf{w}_0$. Since $l(\mathbf{w})$ reaches its minimum value in $\mathbf{w}_0$, we set $\nabla l(\mathbf{w}_0) = 0$ and thus obtain a local approximation of loglikelihood for $\mathbf{w}$:

$$\ln p(\mathbf{y}|\mathbf{X},\ \mathbf{w}) \propto l(\mathbf{w}) - l(\mathbf{w}_0) = \ln \frac{L(\mathbf{w})}{L(\hat{\mathbf{w}})} \approx -\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^{\mathsf{T}}\mathbf{H}(\mathbf{w}-\mathbf{w}_0). \tag{8}$$

Thus $\hat{\mathbf{w}}$ is locally normal

$$\hat{\mathbf{w}} \sim \mathcal{N}(\mathbf{w}_0,\ \mathbf{H}^{-1}) \tag{9}$$

with the covariance matrix $\Sigma = H^{-1}$.

3 Specificity-based outlier detection

In this section we describe the proposed method of outlier detection. Consider an object $(\mathbf{x}_i,\ y_i)$ from the training sample $i \in \mathcal{S}$. If $\hat{\mathbf{w}}$ estimates, given by (4), differ significantly for $\mathcal{S}$ and $\mathcal{S} \setminus \{i\}$, then the object is called an *outlier*. Further we introduce the specificity measure as a statistic for testing parameter difference for significance.

Let $\Delta_i\mathbf{w}$ denote the difference between parameter vectors $\hat{\mathbf{w}}_i$ and $\hat{\mathbf{w}}$

$$\Delta_i\mathbf{w} = \hat{\mathbf{w}}_i - \hat{\mathbf{w}}$$

where estimate $\hat{\mathbf{w}}_i$ is based on the training sample $\mathcal{S}$ without i-th object

$$\hat{\mathbf{w}}_i = \arg\max_{\mathbf{w}\in\mathbb{R}^n} L(\mathbf{w}|\mathbf{X}_{\mathcal{S}\setminus\{i\}},\ \mathbf{y}_{\mathcal{S}\setminus\{i\}}),$$

and $\mathbf{w}$ is estimated with respect to $\mathcal{S}$ according to (4). Further for each object $(\mathbf{x}_i,\ y_i)$, $i \in \mathcal{S}$ we introduce the *specificity* $\mathrm{Sp}(\mathbf{x}_i,\ y_i)$ measure as follows:

$$\mathrm{Sp}(\mathbf{x}_i,\ y_i) = (\Delta_i\mathbf{w})^{\mathsf{T}}\mathbf{H}(\Delta_i\mathbf{w}). \tag{10}$$

and $\hat{\mathbf{w}}$.

Since by (9) the estimates $\hat{\mathbf{w}}$ are locally normal, $\Delta_i\mathbf{w}$ follows normal distribution $\mathcal{N}(\mathbf{0},\ \mathbf{H}^{-1})$ for every $i \in \mathcal{B}$. Therefore, if the object $(\mathbf{x}_i,\ y_i)$ is not an outlier, and $\mathbf{H}$ is not degenerate, then $\mathrm{Sp}(\mathbf{x}_i,\ y_i)$ follows χ^2 distribution with n degrees of freedom:

$$\mathrm{Sp}(\mathbf{x}_i,\ y_i) = (\Delta_i\mathbf{w})^{\mathsf{T}}\mathbf{H}(\Delta_i\mathbf{w}) \sim \chi^2(n). \tag{11}$$

This allows us to use specificity $\mathrm{Sp}(\mathbf{x}_i,\ y_i)$ as a test statistic to detect outliers: given significance level α, the object $(\mathbf{x}_i,\ y_i)$ is considered an outlier if $\mathrm{Sp}(\mathbf{x}_i,\ y_i)$ exceeds the corresponding quantile of $\chi^2(n)$.

Modification of the object selection method. Application of the proposed method is restricted to the cases of invertible $\mathbf{H}$ estimation and must be adapted for the case of degenerate $\mathbf{H}$. In the latter case specificity follows χ^2 distribution with $\mathrm{rg}(\mathbf{H})$ degrees of freedom instead of n as follows from (11). Since $\mathbf{H}$ is computed with some fixed precision, one must specify some threshold $\lambda_0 \geq 0$ such that the eigenvalue λ of $\mathbf{H}$ is set to zero if $\lambda \leq \lambda_0$. Moreover, since $\mathbf{H}$ is degenerate, the procedure (6), which involves $\mathbf{H}$ inversion, is no longer applicable.

A common way to deal with inversion of ill-conditioned matrices is regularization (Neumaier 1998). To regularize a degenerate covariance matrix $\mathbf{H}$ assume a normal $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0,\ \tau\mathbf{I})$ prior distribution $p(\mathbf{w})$ over $\mathbf{w}$ (Li and Goel 2006) with covariance matrix $\tau\mathbf{I}$ for some $\tau > 0$. Using (7) and (8), derive posterior distribution of the estimates $\hat{\mathbf{w}}$:

$$\hat{\mathbf{w}} \sim \mathcal{N}\left(\mathbf{w}_0,\ \left(\mathbf{H} + \tfrac{1}{\tau}\mathbf{I}\right)^{-1}\right).$$

Regularizing hessian matrix $\mathbf{H}$ yields a *regularized specificity*

$$\mathrm{Sp}_\tau(\mathbf{x}_i,\ y_i) = (\Delta_i\mathbf{w})^\mathsf{T}\left(\mathbf{H} + \tfrac{1}{\tau}\mathbf{I}\right)(\Delta_i\mathbf{w}).$$

Note that $\mathrm{Sp}_\tau(\mathbf{x}_i,\ y_i) \to \mathrm{Sp}(\mathbf{x}_i,\ y_i)$ as $\tau \to \infty$.

Introducing a normal prior distribution for $\mathbf{w}$ fixes the issue with the degenerate matrix $\mathbf{H}$ inversion, but requires specifying an additional parameter $\tau \geq 0$. Further we suggest an empirical approximation of specificity, which does not involve any additional parameters. For this purpose we consider the sample estimation

$$D_j = \frac{\sum_{i\in\mathcal{S}}(\Delta_i w_j)^2}{|\mathcal{S}| - 1}.$$

of variance $\mathrm{Var}(w_j)$ for each element of parameter vector $\mathbf{w} = [w_1, \ldots, w_n]$. Using the empirical variance D_j we introduce an approximation of specificity measure

$$\mathrm{Sp}_w(\mathbf{x}_i,\ y_i) = \sum_{j=1}^{n} \frac{(\Delta_i w_j)^2}{D_j}, \tag{12}$$

where $\mathrm{Sp}_w(\mathbf{x}_i, y_i)$ is called the *empirical specificity*. This approximation of $\mathrm{Sp}(\mathbf{x}, y)$ will be used in experiments to implement the proposed method. The empirical specificity $\mathrm{Sp}_w(\mathbf{x},\ y)$ is preferable to $\mathrm{Sp}(\mathbf{x},\ y)$ for object selection as $\mathrm{Sp}_w(\mathbf{x},\ y)$ does not depend on possibly ill-conditioned or even degenerate matrix $\mathbf{H}$. Further we demonstrate that the empirical specificity $\mathrm{Sp}_w(\mathbf{x},\ y)$ induces nearly the same order on the training sample and thus can be used to detect outliers instead of $\mathrm{Sp}(\mathbf{x},\ y)$.

4 Computational experiment

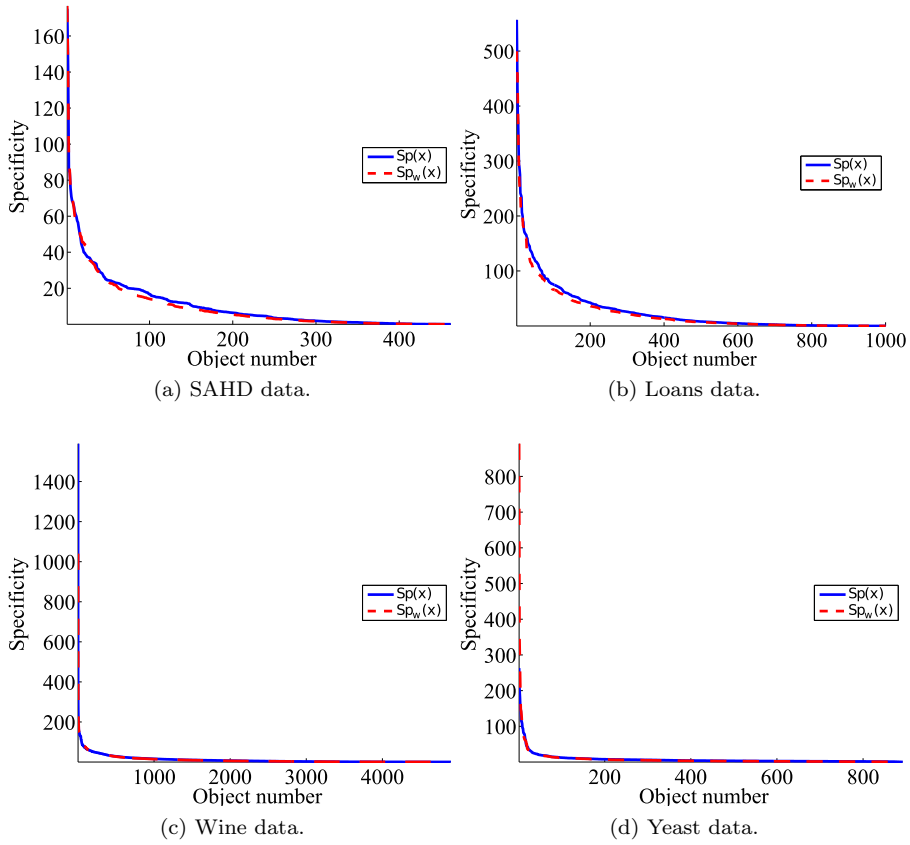
The aim of the computational experiment is to analyze the suggested method for object selection. We use four benchmark datasets from the UCI machine learning repository. Some datasets originally were construsted for multiclass classefication. In such cases we artificially reduced them to binary classification. The details are given below.

1. German cash loans dataset (German dataset 1994) contains 1000 instances, 24 attributes, 2 classes.
2. Heart disease in South Africa dataset (SAHD dataset 1993) contains 462 instances, 13 attributes, 2 classes.
3. Wine quality dataset (Wine data 1991) contains 4898 instances, 11 attributes and 2 classes. For this dataset class labels range from 0 to 10 indicating wine quality. We assigned $y = 0$ class labels to low quality (0–5) wine samples and $y = 1$ class labels to high quality (6–10) wine samples.
4. Yeast dataset (Yeast dataset 1996) has 892 instances, 8 attributes and 2 classes. The dataset contains information about protein localization in cell. We used two biggest classes from this dataset for classification.
5. Housing dataset(Housing dataset 1978) The dataset contains information on house prices in Boston. Houses priced over 25,000$ were labeled as $y = 1$, ones with prices below 25,000$ were labeled as $y = 0$. The dataset contains 506 instances, 13 attributes and 2 classes.
6. Breast cancer dataset (Breast cancer dataset 1992). This dataset contains 699 instances, 9 attributes, 2 classes.
7. Contraceptive dataset (Contraceptive dataset 1987). For this dataset class labels may take 3 values, indicating the usage of contraceptives: 1 for no usage, 2 for long-term usage, and 3 for short-term usage. We assigned $y = 0$ class labels to no usage class and $y = 1$ to both long-term and short-term usage classes. The dataset contains 1473 instances, 9 attributes and 2 classes.
8. Australian cash loans dataset (Australian dataset 1987) contains 690 instances, 14 attributes, and 2 classes.

All the attributes in all cases are numerical. We use logistic regression to solve the classification problem and evaluate classification quality with AUC measure (Ling et al. 2003). Further we prove that quality improvement concerned with removing outliers from the sample D is statistically significant.

Firstly we compare two specificity criteria, $\mathrm{Sp}(\mathbf{x},\, y)$ (10) and $\mathrm{Sp}_w(\mathbf{x},\, y)$ (12). To demonstrate that these measures are equivalent, we show that they induce nearly the same order on the training sample. The solid lines in Figures 1, 2 show the plot of the normalized specificity $\mathrm{Sp}(\mathbf{x},\, y)$ versus object index i, with objects $(\mathbf{x}_i,\, y_i)$ ordered by $\mathrm{Sp}(\mathbf{x},\, y)$ in descending order. Similarly, we sort the objects of the sample D by $\mathrm{Sp}_w(\mathbf{x},\, y)$ in descending order and plot the resulting curve with the dashed 1, 2. Note that for all tested datasets there are few objects with high specificity $\mathrm{Sp}(\mathbf{x},\, y)$ or empirical specificity $\mathrm{Sp}_w(\mathbf{x},\, y)$.

Further we compute Kendall and Pearson rank correlation coefficients between $\mathrm{Sp}(\mathbf{x},\, y)$ and $\mathrm{Sp}_w(\mathbf{x},\, y)$ (see Table 1). For all datasets, except for the yeast dataset, we observed strong positive linear and monotonous connection between empirical specificity $\mathrm{Sp}_w(\mathbf{x},\, y)$ and specificity $\mathrm{Sp}(\mathbf{x},\, y)$: both Kendall and Pearson rank correlation coefficients are close to 1. For yeast dataset, Pearson correlation between $\mathrm{Sp}(\mathbf{x})$ and $\mathrm{Sp}_w(\mathbf{x},\, y)$ is moderate, but Kendall correlation is close to 1. In the experiments with artificial data, which we will discuss later in this Section, we observed correlations between $\mathrm{Sp}(\mathbf{x})$ and $\mathrm{Sp}_w(\mathbf{x},\, y)$ above 0.8 for Pearson correlation and above 0.7 for Kendall correlation even for high contamination fractions. Thus the order induced by empirical specificity $\mathrm{Sp}_w(\mathbf{x},\, y)$ on a sample set is nearly the same as the order induced by $\mathrm{Sp}(\mathbf{x},\, y)$. Since empirical specificity does not involve ill-conditioned (or even degenerate) matrix $\mathbf{H}$, it is preferable to use $\mathrm{Sp}_w(\mathbf{x},\, y)$ for



(a) SAHD data. (b) Loans data.

(c) Wine data. (d) Yeast data.

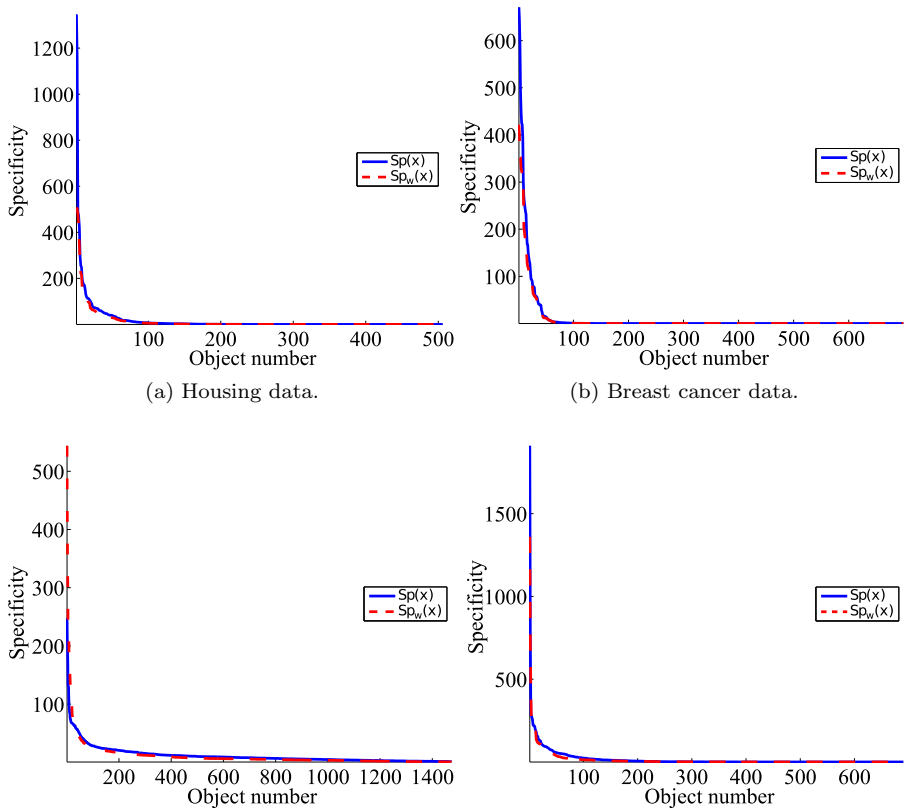
Fig. 1: Comparison of specificity functions $\mathrm{Sp}(\mathbf{x})$ and $\mathrm{Sp}_w(\mathbf{x})$.

object selection, and we will use it in our experiments to detect outliers. To esti-

Table 1: Correlations of specificities $\mathrm{Sp}(\mathbf{x},\, y)$ (10) and $\mathrm{Sp}_w(\mathbf{x},\, y)$ (12).

Data \ Correlations	Pearson	Kendall
SAHD	0.9736	0.9132
Loans	0.9794	0.9377
Wine	0.9528	0.9028
Yeast	0.5230	0.8597
Housing	0.8903	0.91505
Breast cancer	0.9657	0.9760
Contraceptive	0.9291	0.8819
Australian loans	0.9763	0.9343

mate the quality of the proposed method we train logistic model (1) two times —



(a) Housing data. (b) Breast cancer data.

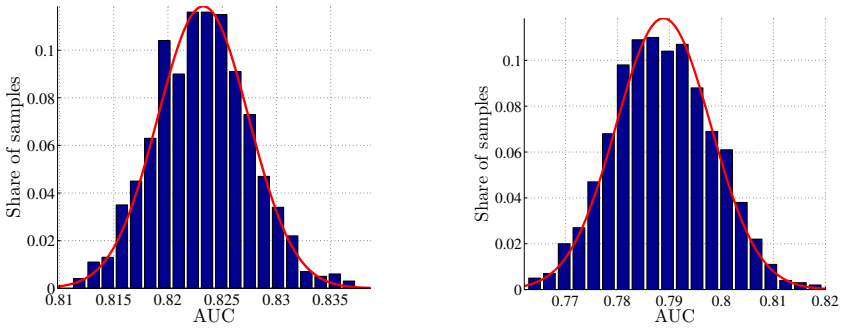
(c) Contraceptive data. (d) Australian loans data.

Fig. 2: Comparison of specificity functions $\mathrm{Sp}(\mathbf{x})$ and $\mathrm{Sp}_w(\mathbf{x})$.

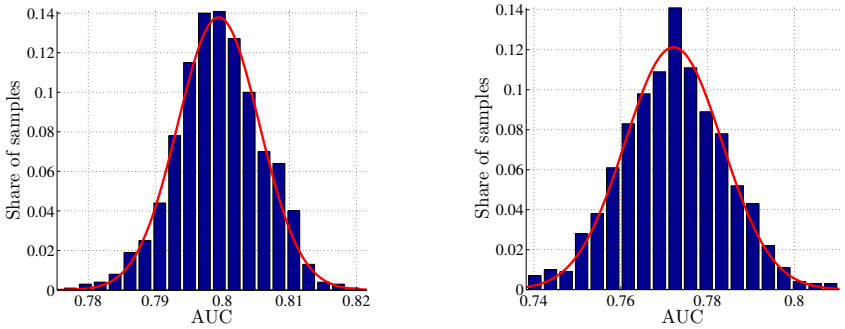
before and after removing outliers — and compute corresponding AUC measures. Table 2 shows the increase of AUC after outlier extraction for both samples.

Table 2: Change in AUC caused by outlier filtering based on the proposed object selection procedure.

Data	AUC before selection	AUC after selection	# of deleted objects
SAHD	0.7948	0.8275	15 out of 462
Loans	0.8179	0.8779	50 out of 1000
Wine	0.7992	0.8105	48 out of 4898
Yeast	0.7123	0.7332	18 out of 892
Housing	0.9585	0.9891	14 out of 506
Breast cancer	0.9943	0.9993	15 out of 699
Contraceptive	0.6799	0.7023	33 out of 1473
Australian loans	0.9351	0.9638	22 out of 690



(a) Loans data. Training. (b) Loans data. Testing.



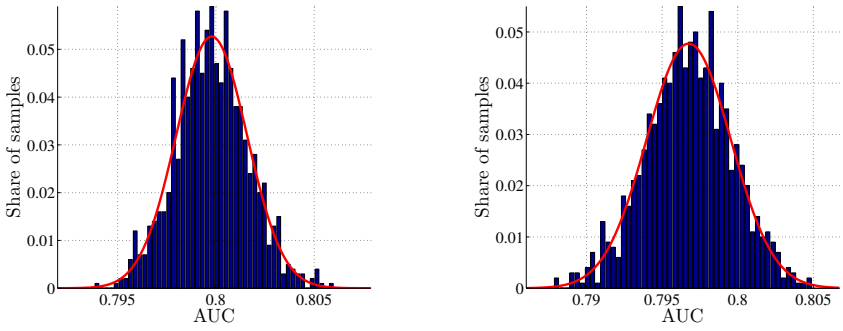
(c) SAHD data. Training. (d) SAHD data. Testing.

Fig. 3: Empirical distribution of AUC and its normal approximation for loans and heart disease data.

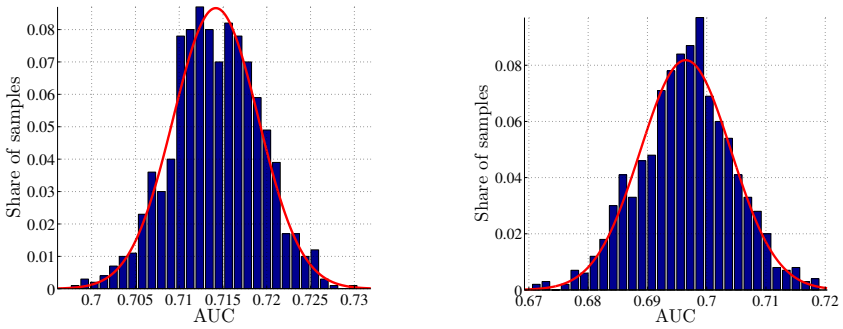
Since classification quality depends on both training $(\mathbf{X}_{\mathcal{S}}, \mathbf{y}_{\mathcal{S}})$ and test $(\mathbf{X}_{\mathcal{T}}, \mathbf{y}_{\mathcal{T}})$ samples, the increase of AUC is possible even after removing several objects from the sample at random. To prove that AUC improvement after removing outliers is statistically significant we test the hypothesis H_0 that the improvement is caused only by reduction of the sample set size.

To test H_0 hypothesis, we estimate its significance level p through the following procedure. We sample at random 1000 subsets of a smaller size from each tested dataset. Denote the j-th generated subset by D_j. We randomly split D_j into training and testing samples 50 times. Table 3 presents subsampling information and training sizes for each dataset. For each subset D_j we estimate model parameters $\hat{\mathbf{w}}^j$ using procedure (2) and compute the corresponding AUC(j) value. Thus we obtain a number of observations of AUC values for both training and test samples. We use them to compute the empirical distribution of AUC. Then an estimate of p is the share of samples D_j with AUC(j) value higher than AUC reached after objects selection. Since even for 1000 generated samples there is no D_j with AUC(j) higher than AUC obtained after object selection, this definition yields $p = 0$.

To obtain non-zero observed significance level p, we use normal approximation of the empirical distribution of AUC and test empirical distribution for normality using Shapiro and Wilk test. Figures 3– 6 show histograms for the empirical dis-



(a) Wine data. Training. (b) Wine data. Testing.



(c) Yeast data. Training. (d) Yeast data. Testing.

Fig. 4: Empirical distribution of AUC and its normal approximation for wine and yeast data.

Table 3: Initial sample size, size of reduced sample and training sample for each dataset.

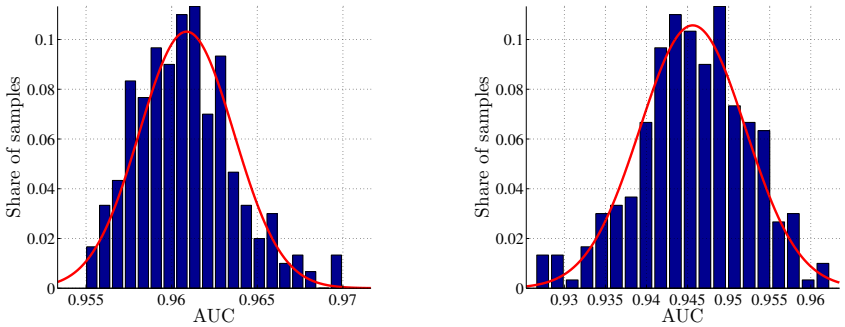
Data	Sample size	Resampling size	Training size
SAHD	462	447	300
Loans	1000	950	690
Wine	4898	4840	3000
Yeast	892	874	550
Housing	506	492	350
Breast cancer	699	684	50
Contraceptive	1473	1440	800
Australian loans	690	668	400

tributions of AUC(j), $j = 1, \ldots, 1000$ for training and testing samples and their normal approximations.

The properties of empirical distributions and their normal approximations are summarized in Table 4. AUC values given in the first row of the table reflects classification quality for training and testing samples after removing outliers. Further rows of Table 4 present statistical properties of the sampled AUC(j), $j = 1, \ldots, 1000$ for training and testing samples. For all considered datasets apart

Table 4: Empirical distributions of AUC and their normal approximations.

	Loans		SAHD	
Properties	Train	Test	Train	Test
AUC value	0.8819	0.8308	0.8507	0.8093
AUC expectation estimate, $\hat{m}$	0.8233	0.7889	0.7994	0.7722
AUC variance estimate, $\hat{\sigma}^2$	$1.75 \cdot 10^{-5}$	$8.3 \cdot 10^{-5}$	$3.7 \cdot 10^{-5}$	$1.3 \cdot 10^{-4}$
AUC standard deviation estimate, $\hat{\sigma}$	0.0042	0.0091	0.0061	0.011
$\lvert \text{AUC} - \hat{m} \rvert / \hat{\sigma}$	14.0	6.8	5.15	3.32
Observed p-value for Shapiro-Wilk test, p_{SW}	0.2655	0.2364	0.2786	0.7879
Observed signif. level for H_0, p_0	0	$5.3 \cdot 10^{-12}$	$1.3 \cdot 10^{-7}$	$4.6 \cdot 10^{-4}$
	Wine		Yeast	
Properties	Train	Test	Train	Test
AUC value	0.8109	0.8084	0.7346	0.7225
AUC expectation estimate, $\hat{m}$	0.7998	0.7968	0.7142	0.6965
AUC variance estimate, $\hat{\sigma}^2$	$3.26 \cdot 10^{-6}$	$7.8 \cdot 10^{-6}$	$2.4 \cdot 10^{-5}$	$5.8 \cdot 10^{-5}$
AUC standard deviation estimate, $\hat{\sigma}$	0.0018	0.0028	0.0049	0.0076
$\lvert \text{AUC} - \hat{m} \rvert / \hat{\sigma}$	6.15	4.15	4.18	3.41
Observed p-value for Shapiro-Wilk test, p_{SW}	0.3103	0.6989	0.5326	0.4288
Observed signif. level for H_0, p_0	$3.9 \cdot 10^{-10}$	$1.7 \cdot 10^{-5}$	$1.4 \cdot 10^{-5}$	$3.2 \cdot 10^{-4}$
	Housing		Breast cancer	
Properties	Train	Test	Train	Test
AUC value	0.9900	0.9826	0.9999	0.9976
AUC expectation estimate, $\hat{m}$	0.9609	0.9456	0.9991	0.9856
AUC variance estimate, $\hat{\sigma}^2$	$7.6 \cdot 10^{-6}$	$4.2 \cdot 10^{-5}$	$9.4 \cdot 10^{-7}$	$1.1 \cdot 10^{-5}$
AUC standard deviation estimate, $\hat{\sigma}$	0.0028	0.0064	0.00097	0.0033
$\lvert \text{AUC} - \hat{m} \rvert / \hat{\sigma}$	10.55	5.75	0.82	3.62
Observed p-value for Shapiro-Wilk test, p_{SW}	0.0834	0.1199	0	0
Observed signif. level for H_0, p_0	0	$4.6 \cdot 10^{-9}$	0.2050	$1.5 \cdot 10^{-4}$
	Contraceptive		Australian loans	
Properties	Train	Test	Train	Test
AUC value	0.7048	0.6881	0.9647	0.9567
AUC expectation estimate, $\hat{m}$	0.6831	0.6633	0.9372	0.9207
AUC variance estimate, $\hat{\sigma}^2$	$2.28 \cdot 10^{-5}$	$3.35 \cdot 10^{-5}$	$1.0 \cdot 10^{-5}$	$2.5 \cdot 10^{-5}$
AUC standard deviation estimate, $\hat{\sigma}$	0.0048	0.0058	0.0032	0.0050
$\lvert \text{AUC} - \hat{m} \rvert / \hat{\sigma}$	4.56	4.29	8.67	7.25
Observed p-value for Shapiro-Wilk test, p_{SW}	0.1596	0.7949	0.3047	0.3966
Observed signif. level for H_0, p_0	$2.6 \cdot 10^{-6}$	$8.8 \cdot 10^{-6}$	0	$2.1 \cdot 10^{-13}$



(a) Housing data. Training.

(b) Housing data. Testing.

(c) Breast cancer data. Training.

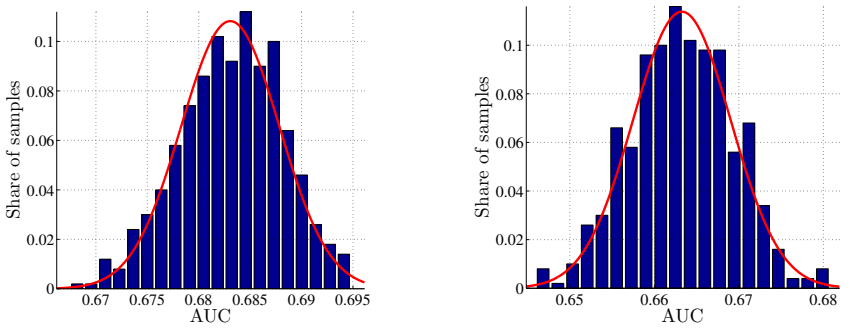
(d) Breast cancer data. Testing.

Fig. 5: Empirical distribution of AUC and its normal approximation for housing and breast cancer data.

from breast cancer dataset the observed significance level for H_0 is less than 10^{-3} both for training and testing samples. This corresponds to more than 3 standard deviations from the mean value. Therefore, the hypothesis H_0 is rejected. Thus we conclude that object selection has statistically significant effect in terms of AUC value. For breast cancer dataset this does not hold for training sample because the value of AUC reached before outlier filtering is already close to the maximum possible value of 1. The same reason causes the deviation from normality for this dataset as shown by low p-value for Shapiro-Wilk test.

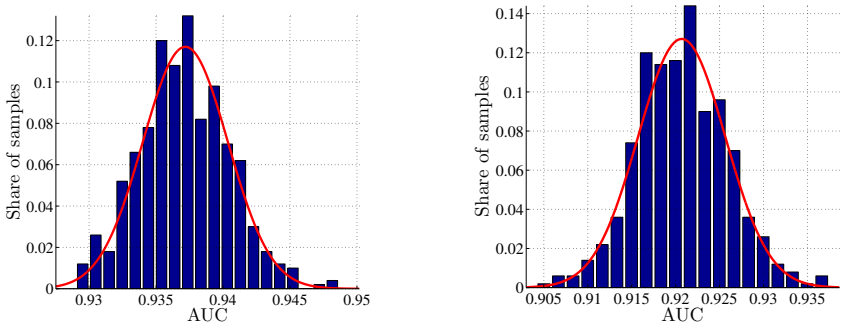
Filtering clustered and non-clustered outliers. In this paragraph we study how the proposed algorithm performs in application to artificial datasets. To generate artificial datasets we consider two ways of modelling outliers, namely clustered and non-clustered outliers. Non-clustered outliers are individual atypical objects. Detecting such outliers may be seen as a problem of one-class classification. Clustered outliers share some common properties and may be seen as objects from another class. Using the AUC value of 0.7 as the threshold for acceptable discrimination (Hosmer et al. 2000), we define the maximum contamination fraction (share of outliers) for which the AUC is still above the threshold.

Artificial data with non-clustered outliers is generated as follows. Features $\mathbf{x}_i$ are sampled from normal distribution $N(\mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is 2×2 identity matrix.



(a) Contraceptive data. Training.

(b) Contraceptive data. Testing.



(c) Australian loans data. Training.

(d) Australian loans data. Testing.

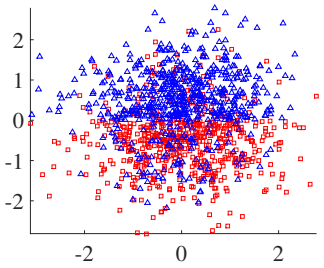
Fig. 6: Empirical distribution of AUC and its normal approximation for contraceptive and australian loans data.

Objects are labeled with $y_i = 1$ for $\mathbf{x}_i$ if $x_2 > 0$ and $y_i = 0$ otherwise. Outliers are sampled from the same distribution but labeled using the opposite rule: $y_i = 0$ for $\mathbf{x}_i$ if $x_2 > 0$ and $y_i = 1$ otherwise. Figure 7a shows the generated sample for 1000 regular objects and 200 outliers.

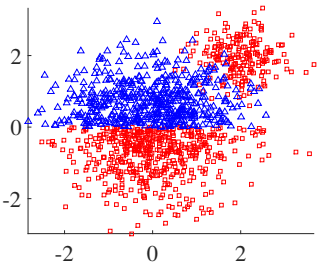
Artificial data with clustered outliers is generated as follows. Non-outliers are generated from $N(\mathbf{0},\, \mathbf{I})$, where $\mathbf{I}$ is 2×2 identity matrix. For non-outliers we assign $y_i = 1$ for $\mathbf{x}_i$ if $x_2 > 0$ and $y_i = 0$ otherwise. Outliers are generated from $N([2,\, 2]^\intercal,\, 0.5\mathbf{I})$, where $\mathbf{I}$ is 2×2 identity matrix. All outliers have the class label of 0. Figure 7b shows the generated sample for 1000 regular objects and 200 outliers.

We examined the proposed method of outlier detection for various contamination fractions from 0 to 50%. Figure 8 shows the dependence of AUC after removing outliers on contamination fractions.

For non-clustered outliers the threshold of 0.7 is reached for contamination fraction of 41.1%. For clustered outliers the same threshold is reached for contamination fraction of 33.3%. These results show that the method is applicable even for datasets with high shares of outliers. However, the method performs better for datasets with non-clustered outliers. Note that for both clustered and non-clustered outliers for all considered contamination fractions the correlation between specificity and empirical specificity is above 0.8 for Pearson correlation and above 0.7 for Kendall correlation. Thus even for highly-contaminated datasets

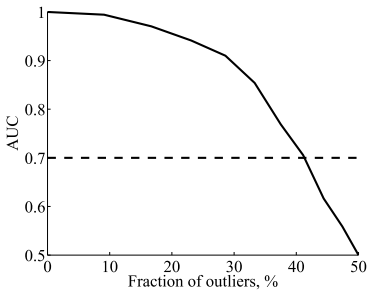


(a) Artificial sample. Non-clustered outliers.

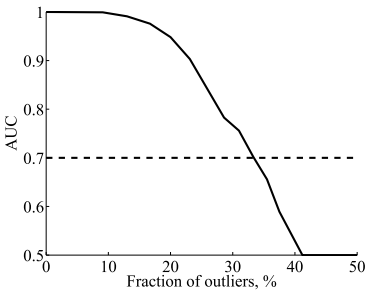


(b) Artificial sample. Clustered outliers.

Fig. 7: Artificial datasets having clustered and non-clustered outliers.



(a) AUC dependence on contamination fraction. Non-clustered outliers.



(b) AUC dependence on contamination fraction. Clustered outliers.

Fig. 8: Dependence of AUC on contamination fraction.

empirical specificity (12) can be used instead of specificity (10) in case of ill-conditioned hessian matrix $\mathbf{H}$.

Comparison to other methods. In this paragraph we compare introduced specificity measure with four other widely used for outliers filtering measures: bayesian, Pearson, deviance residuals (Albert and Chib 1995) and gamma plots based measure (Evans and Jones 2002). However, though Pearson and deviance residuals are defined differently, they induce the same order on objects and therefore give the same results for outliers filtering.

We compare the proposed method to alternatives using cross validation. The sample is splitted into learning and testing samples. We use apply each method to detect outliers and remove them from the learning sample. Each time we use filtered learning sample to train the model and classify the testing sample. We compare methods measuring classification quality in terms of AUC.

For comparison we use UCI benchmark datasets as well as artificial datasets described above. For each dataset we split it into learning and testing sample 100 times. Table 5 lists AUC values for the filtered datasets and the results of testing AUC difference for significance for different methods. Column specificity corresponds to the results obtained with suggested method for outliers filtering

Table 5: Comparison of outlier detection methods.

UCI data	Pearson / deviance	Bayes	Gamma	Specificity	t_p	t_b	t_g
SAHD	0.7716	0.7676	**0.7722**	0.7661	-1.6395	-0.448	-1.818
Loans	**0.7868**	0.7864	0.7828	0.7802	-2.7093	-2.5345	-1.0673
Wine	**0.7977**	0.7974	0.7972	0.7970	-0.8471	-0.4220	-0.2471
Yeast	0.6845	**0.6951**	0.6937	0.6944	5.8773	-0.3997	0.4233
Housing	0.9420	0.9435	0.9388	**0.9439**	11.4120	2.1089	30.6347
Breast cancer	**0.9923**	0.9917	0.9921	0.9923	-0.7188	15.1958	2.7909
Contraceptive	**0.6609**	0.6595	0.6564	0.6590	-1.5758	-0.4427	2.2328
Australian loans	0.9116	0.9123	**0.9172**	0.9145	4.0912	3.2093	-3.8142
Artificial data	Pearson / deviance	Bayes	Gamma	Specificity	t_p	t_b	t_g
Non-clustered, 9.1%	0.8997	0.9021	**0.9031**	0.9002	0.2450	-1.1300	-1.7346
Non-clustered*, 9.1%	0.8945	0.8956	0.8956	**0.8958**	0.8014	0.1583	0.1324
Non-clustered, 23.1%	0.7646	0.7653	**0.7885**	0.7665	0.7945	0.5036	-10.9947
Non-clustered*, 23.1%	0.7671	0.7593	0.7692	**0.7694**	0.9949	4.3273	0.0926
Non-clustered, 33.3%	0.6673	0.6679	0.6680	**0.6680**	0.6450	0.1075	0.0305
Non-clustered*, 33.3%	0.5372	0.6666	0.5817	**0.6681**	64.5832	0.7482	42.6279
Clustered, 9.1%	0.8885	0.9261	0.8673	**0.9269**	20.9410	0.4443	32.5022
Clustered*, 9.1%	0.8740	0.9515	0.8877	**0.9541**	66.9012	2.1318	55.4587
Clustered, 16.7%	0.8393	**0.8471**	0.8275	0.8456	2.5400	-0.6264	7.2975
Clustered*, 16.7%	0.8379	0.8305	0.8112	**0.9060**	44.4005	49.1751	61.7457
Clustered, 23.1%	0.8107	0.8171	0.7975	**0.8174**	3.4906	0.1210	10.3676
Clustered*, 23.1%	0.8105	0.7923	0.7945	**0.8113**	0.2828	6.5297	5.7736
Clustered, 33.3%	0.7860	0.7856	**0.7872**	0.7853	-0.4075	-0.1803	-1.1061
Clustered*, 33.3%	0.7675	**0.7762**	0.7713	0.7671	-0.1078	-2.4158	-1.1150

while columns Pearson/deviance, Bayes and Gamma correspond to the results obtained with competitive methods.

Denote by AUC_p, AUC_b, AUC_g and AUC_s values of AUC for testing sample observed using Pearson (deviance) residuals, bayesian residuals, gamma plot method and specificity to filter outliers in learning sample respectively. Student's t-statistics for significance of AUC_s - AUC_p, AUC_s - AUC_b and AUC_s - AUC_g correspondingly are denoted by t_p, t_b and t_g.

Table 5 lists the results of comparison. For each artificial dataset the table indicates whether it has clustered or non-clustered outliers and its contamination fraction. Asterisks mark the cases where a share of removed objects exceeded the actual contamination fraction for all methods. Table 5 shows similar results for all 8 UCI benchmark datasets. The suggested method is the best one only for one dataset out of eight considered ones. However, as the values t_p, t_b and t_g demonstrate, most of the differences between methods are insignificant. If we consider only the cases where the differences are significant ($|t| > 2$), we find that the suggested method outperforms its alternatives on one versus one basis for 3 benchmark dataset while being significantly worse only for a single benchmark dataset.

For artificial datasets the proposed method performs generally better for both clustered and non-clustered outliers. For non-clustered outliers suggested method and method based on gamma plots show similar performance while for clustered outliers the latter method works much worse than the suggested one. The values of t_p, t_b and t_g are especially high in the case where regular objects were detected as outliers. One possible explanation for this is that both pearson/deviance residuals and bayesian residuals recognise objects as outliers if they are poorly described by the model. This property can be beneficial if the dataset has a small number of outliers (Kosinski 1998). However, for datasets with high contamination fraction such methods are not effective. Specificity can be more effective because it estimates the impact of each object on the model stability instead of measuring how well the object fits the model.

Conclusion

In this paper we consider the problem of object selection in banking credit scorecards construction. To design a reliable banking credit scorecard one should select an informative set of objects for the training set. This paper proposes a new method for outlier detection. The parametric scoring model is constructed as a logistic regression model using the selected sample set. The method is based on the newly introduced specificity measure. Specificity measures the impact of each object of the sample on regression parameters. We used the fact that specificity follows χ^2 distribution for non-outliers to perform filtering. To adapt the proposed method for the common case of ill-conditioned hessian matrix we introduce the empirical specificity, which does not involve inversion of hessian matrix. Computational experiments show high positive and monotonous correlation between specificity and empirical specificity.

We compare our method to Pearson residuals, bayesian residuals, and gamma-plots using rel and artificially generated data. For real datasets we find that the proposed method wins over the alternatives on one-versus-one basis and performs generally better for artificial datasets with both clustered and non-clustered outliers.

We observe reasonable quality for artificial datasets with up to 40% of non-clustered outliers and 30% of clustered outliers. To process sample sets with larger clusters of outliers one may use multilevel model or mixture of models. This is the subject to the further research.

References

Albert and Chib 1995. Albert J., Chib S. (1995) Bayesian residual analysis for binary response regression models // Biometrika, 82(4), 747–769.

Australian dataset 1987. Australian credit approval data. URL: http://archive.ics.uci.edu/ml/datasets/Statlog+%28Australian+Credit+Approval%29/ (1987). Last checked: 29.01.2016.

Bishop 2006. Bishop C. M. Pattern recognition and machine learning. Springer, 2006.

Bishop and Nasrabadi 2006. Bishop C. M., Nasrabadi N. M. Pattern recognition and machine learning. // Journal of electronic imaging, 2007. Vol. 16. No. 4.

Boyd and Vandenberghe 2004. S. Boyd and L. Vandenberghe (2004). Convex Optimization. Cambridge: Cambridge University Press.

Breast cancer dataset 1992. Breast cancer data. URL: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29/ (1992). Last checked: 29.01.2016.

Contraceptive dataset 1987. Contraceptive method choice data. URL: http://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice/ (1987). Last checked: 29.01.2016.

Cook and Weisberg 1989. Cook, R.D. and Weisberg, S., "Residuals and Influence in regression," London : Chapman & Hall, 1989.

Croux and Haesbroeck 2003. Croux C., Haesbroeck G. (2003). Implementing the Bianco and Yohai estimator for Logistic Regression // Computational Statistics and Data Analysis, 44, 273–295.

Evans and Jones 2002. Evans D., and Jones A. (2002). A proof of the Gamma test // Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 458(2027), 2759-2799.

Gelman et al. 2003. A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin (2003). Bayesian Data Analysis. Chapman & Hall.

German dataset 1994. German cash loan data. URL: http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/ (1994). Last checked: 04.05.2014.

Filzmoser et al. 2008. Filzmoser, P., Maronna, R. and Werner, M. (2008). Outlier identification in high dimensions. Computational Statistics & Data Analysis 52.3, 1694-1711.

Hahn and Soyer 2005. Hahn E. D., and Soyer R. (2005) Probit and logit models: Differences in the multivariate realm. Submitted to The Journal of the Royal Statistical Society, Series B.

Hardin and Hilbe 2007. Hardin J. W., and Hilbe J. M.. (2007). Generalized linear models and extensions. Stata Press.

Hardin and Rocke 2004. Hardin, J., and Rocke, D. M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Computational Statistics & Data Analysis 44.4, 625-638.

Hosmer et al. 2000. (2013). Hosmer D. W., Lemeshow S., Sturdivant R. X. Applied logistic regression. Wiley. com.

Housing dataset 1978. Housing data. URL: http://archive.ics.uci.edu/ml/datasets/Housing/ (1978). Last checked: 29.01.2016.

Kosinski 1998. Kosinski, Andrzej S. (1998). A procedure for the detection of multivariate outliers." Computational statistics & data analysis 29.2, 145-161.

Li and Goel 2006. Bin Li and Prem K. Goel (2006). Regularized optimization in statistical learning: a bayesian perspective
Statistica Sinica 16: 411-424.

Ling et al. 2003. *Ling C. X., Huang J., Zhang H.* (2003). AUC: a statistically consistent and more discriminating measure than accuracy // International joint Conference on artificial intelligence, 18, 519–526.

Malkovich and Afifi 1973. *Malkovich J. F., Afifi A. A.* (1973) On tests for multivariate normality // Journal of the American Statistical Association, 68(341), 176–179.

Motrenko et al. 2014. *Motrenko A., Strijov V., Weber G.-W.* (2014). Bayesian sample size estimation for logistic regression // Journal of Computational and Applied Mathematics, 255, 743-752.

Neumaier 1998. A. Neumaier (1998). Solving ill-conditioned and singular linear systems: A tutorial on regularization
SIAM Review 40: 636-666.

Rousseeuw and Zomeren 1990. Rousseeuw, P.J. and Zomeren, B.C.V., (1990). Unmasking Multivariate Outliers and Leverage Points, Journal of the American Statistical Association 85, 633-639.

SAHD dataset 1993. South Africa heart disease data. URL: http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart/ (1993). Last checked: 04.05.2014.

Sebert et al. 1998. Sebert, David M., Douglas C. Montgomery, and Dwayne A. Rollier. (1998). A clustering algorithm for identifying multiple outliers in linear regression. Computational statistics & data analysis 27.4, 461-484.

Siddiqi 2006. Siddiqi N. Credit risk scorecards: developing and implementing intelligent credit scoring. Wiley, 2006.

Wine data 1991. Wine quality data. URL: http://archive.ics.uci.edu/ml/datasets/Wine/ (1991). Last checked: 04.05.2014.

Wisnowskia et al. 2001. Wisnowski, James W., Douglas C. Montgomery, and James R. Simpson. (2001). A comparative analysis of multiple outlier detection procedures in the linear regression model. Computational statistics & data analysis 36.3: 351-382.

Yeast dataset 1996. Yeast data. URL: http://archive.ics.uci.edu/ml/datasets/Yeast/ (1996). Last checked: 04.05.2014.