www.sjm06.com

Serbian Journal of Management 8 (1) (2013) 9 - 24

Serbian Journal of Management

HIGHLY ROBUST METHODS IN DATA MINING

Jan Kalina*

Institute of Computer Science of the Academy of Sciences of the Czech Republic,
Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic

(Received 14 January 2013; accepted 4 February 2013)

Abstract

This paper is devoted to highly robust methods for information extraction from data, with a special attention paid to methods suitable for management applications. The sensitivity of available data mining methods to the presence of outlying measurements in the observed data is discussed as a major drawback of available data mining methods. The paper proposes several newhighly robust methods for data mining, which are based on the idea of implicit weighting of individual data values. Particularly it propose a novel robust method of hierarchical cluster analysis, which is a popular data mining method of unsupervised learning. Further, a robust method for estimating parameters in the logistic regression was proposed. This idea is extended to a robust multinomial logistic classification analysis. Finally, the sensitivity of neural networks to the presence of noise and outlying measurements in the data was discussed. The method for robust training of neural networks for the task of function approximation, which has the form of a robust estimator in nonlinear regression, was proposed.

Keywords: Data mining, robust statistics, High-dimensional data, Cluster analysis, Logistic regression, Neuralnetworks.

1. STATISTICS VS. DATA MINING

Data mining can be characterized as a process of information or knowledge extraction from data sets, which leads to revealing and investigating systematic associations among invidivual variables. It contains the exploratory data analysis, descriptive modeling, classification, and regression, while classification together with

[*] *Corresponding author: kalina@cs.cas.cz*

DOI: 10.5937/sjm8-3226

regression are commonly denoted as predictive data mining. The concept of data mining can be described as the analytical part of the overall process of extracting useful information from data, which is traditionally called knowledge discovery (Fayyad et al. , 1996; Nisbet et al. , 2009; Martinez et al. , 2011). We can say that data mining has the aim to extract information, while knowledge discovery goes further and aims at acquiring knowledge relevant for the field of expertise; compare Fernandez (2003) and Zvárová et al. (2009). Actually Soda et al. (2010) required data mining methods to be dynamically integrated within knowledge discovery approaches.

In management applications, data mining is often performed with the aim to extract information relevant for making predictions and/or decision making, which can be described as selecting an activity or series of activities among several alternatives. Decision making integrates uncertainty as one of the aspects with an influence on the outcome. The development of computer technology has allowed to implement partially or fully automatic decision support systems, which can be described as very complicated systems offering assistance with the decision making process with the ability to compare different possibilities in terms of their risk. The systems are capable to solve a variety of complex tasks, to analyze different information components, to extract information of different types, and deduce conclusions for management,financial, or econometricapplications (Gunasekaran & Ngai, 2012; Brandl et al. , 2006), allowing to find best available decision within the framework of the evidence-based management (Briner, 2009).

Unlike classical statistical procedures, the data mining methodology does not have the ambition to generalize its results beyond the data summarized in a given database. In data mining, one does not usually assume a random sample from a certain population and the data are analyzed and interpreted as if they constituted the whole population. On the other hand, the work of a statistician is often based on survey sampling, which requires also to propose questionnares, train questioners, work with databases, aggregate the data, compute descriptive statistics, or estimating non-response. Moreover, a statistician commonly deals with observing a smaller number of variables on relatively small samples, which is another difference from the data mining context.

Removal of outliers is one of important steps of validating the plausibility of a model both in statistics and data mining. A common disadvantage of popular data mining methods (linear regression, classification analysis, machine learning) is namely their high sensitivity (non-robustness) to the presence of outlying measurements in the data. Moreover, statisticians have developed the robuststatistical methodology as an alternative approach to some standard procedures, which possess a robustness (insensitivity) to the presence of outliers as well as to standard distributional assumptions. Although the concept of robust data mining describing a methodology for data mining based on robust statistics has been introduced (Shin et al. , 2007), robust methods have not found their way to real data mining applications yet.

This paper proposes new highly robust methods suitable for the analysis of data in management applications. We use idea of implicit weighting to derive new alternative multivariate statistical methods, which ensuresa high breakdown point. This paper has the following structure. Section 2 recalls the highly robust least weighted squares method for estimating parameters in a linear regression models. Section 3 proposes a robust cluster analysis method. Section 4 proposes a robust estimation method for logistic regression, which is used to define a robust multinomial logistic classification in Section 5. Section 6 is devoted to robustness aspects of neural networks and finally Section 7 concludes the paper.

2. LEAST WEIGHTED SQUARES REGRESSION

Because classical statistical methods suffer from the presence of outlying data values (outliers), robust statistical methods have been developed as an alternative approach to data modeling. They originated in 1960sas a diagnostic tool for classical methods (Stigler, 2010), but have developed to reliable self-standing procedures point tailor-made to suppress the effect of data contamination by various kinds of outliers (Hekimoglu et al. , 2009). This section describes the least weighted squares estimator, which is one of promising robust estimators of parameters in the linear regression model. Its idea will be exploited in the following sections.

M-estimators represent the most widely used robust statistical methods (Maronna et al. , 2006). However, as the concept of breakdown point has become the most important statistical measure of resistance (insensitivity) against noise or outlying measurements in the data (Davies & Gather, 2005), the M-estimators have been criticized for their low breakdown point in linear regression (Salibián-Barrera, 2006). Only recently, highly robust methods defined as methods with a high breakdown have been proposed.

The least weighted squares estimator (LWS) proposed by Víšek (2001) is a highly robust statistical tool based on the idea of down-weighting less reliable observations. This estimator is based on implicit weighting of individual observations. The idea is to assign smaller weights to less reliable observations, without the necessity to specify precisely which observations are outliers and which are not.

In the original proposal, the user has to select the magnitudes of the weights. These are assigned to particular observations only after a permutation, which is automatically determined during the computation of the estimator. However, more recent versions of the LWS do not require the user to choose the magnitudes of the weights. Just recently, Čížek (2011) proposed several adaptive versions of the LWS estimator, including a two-stage procedure with data-dependent quantile-based weights. The estimator has a high breakdown point and at the same time a 100 % asymptotic efficiency of the least squares under Gaussian errors. Its relative efficiency is high (over 85 %) compared to maximum likelihood estimators also under various distributional models for moderate samples, as evaluated in a numerical study (Čížek, 2011). Besides, compared to M-estimators, the LWS estimator of regression parameters does not require a simultaneous estimation of the error variance, which is a disadvantage of M-estimators losing a computational simplicity.

The LWS estimator is a weighted analogy of the least trimmed squares (LTS) estimator of Rousseeuw and van Driessen (2006), who considered weights equal to 1 or 0 only. However, the LTS estimator suffers from a high sensitivity to small deviations near the center of the data, while the LWS estimator possesses also a reasonable local robustness(Víšek, 2001). The computation of the LWS estimator is intensive and an approximative algorithm can be obtained as a weighted version of the LTS algorithm (Rousseuw and van Driessen, 2006). Diagnostic tools for the LWS estimator in the linear regression context (mainly for econometric applications) were proposed by Kalina (2011) andhypothesis tests concerning the significance of LWS estimators by Kalina (2012a). Econometric applications of robust statistical methods were summarized by Kalina (2012b). A generalization of the LWS estimator for nonlinear regression will be proposed in Section 6. 4.

3. ROBUST CLUSTER ANALYSIS

Cluster analysis (clustering) is a commonly used data mining tool in management applications. Examples of using the cluster analysis include acustomer segmentation (differentiation) based on sociodemographic characteristics, life style, size of consumption, experience with particular products, etc. (Mura, 2012). Another application is the task of market segmentationallowingto position products or to categorize managers on the basis of their style and strategy (Punj & Stewart, 1983; Liang, 2005). All such tasks solved by grouping units into natural clusters at the explorative stage of information extraction are vital for a successful marketing or management strategy.

Cluster analysis is a general methodology aiming at extracting knowledge about the multivariate structure of given data. It solves the task of unsupervised learning by dividing the data set to several subsets (clusters) without using a prior knowledge about the group membership of each observation. It is often used as an exploratory technique and can be also interpreted as a technique for a dimension reduction of complex multivariate data. Cluster analysis assumes the data to be fixed (non-random) without the ambition for a statistical inference. We can say that it contains a wide variety of methods with numerous possibilities for choosing different parameters and adjusting the whole computation process.

The most common approaches to clustering include the agglomerative hierarchical clustering and *k*-means clustering, which both suffer from the presence of outliers in the data and strongly depend on the choice of the particular method. Some approaches are also sensitive to the initialization of the random algorithm (Vintr et al. , 2012). Robust versions of the clustering have been recently studied in the context of molecular genetics, but they have not penetrated to management applications yet. A robust hierarchical clustering was proposed by García-Escudero et al. (2009) and a robust version of *k*-means cluster analysis tailor-made for high-dimensional applications by Gao and Hitchcock (2010). It is also possible but tedious to identify and manually remove outlying measurements from multivariate data (Svozil, 2008).

In this paper, we propose a new robust agglomerative hierarchical cluster analysis based on a robust correlation coefficient as a measure of similarity between two clusters. The Pearson correlation coefficient is a common measure of similarity between two clusters (Chae et al. , 2008). We propose a robust measure of distance between two clusters based on implicit weighting (Kalina, 2012a) inspired by the LWS regression estimator.

Algorithm 1: (Robust distance between two clusters).

Let us assume two disjoint clusters C_1 and C_2 of *p*-variate observations. Thedistance $d_{LWS}\ (C_1,C_2)$ between the clusters C_1 and C_2 is defined by the following algorithm.

1. Select an observation $X=(X_1,...,X_p)^T$ from C_1 and select an observation $Y=(Y_1,...,Y_p)^T$ from C_2.

2. Consider the linear regression model

$$Y_i=\beta_0 + \beta_1 X_i + e_i, \quad i=1,...p. \tag{1}$$

3. Compute the LWS estimator in the model (1). The optimal weights determined by the LWS estimator will be denoted by $w^*=(w_1,. . . ,w_p)^T$.

4. The LWS-distance between X and Y denoted by $d_{LWS}\ (X,Y)$ will be defined as

$$d_{LWS}(X,Y)=\frac{\sum_{i=1}^{p} w_i^*(X_i-\overline{X}_w)(Y_i-\overline{Y}_w)}{\sqrt{\sum_{j=1}^{p}\left[w_j^*(X_j-\overline{X}_w)\right]^2 \sum_{j=1}^{p}\left[w_j^*(Y_j-\overline{Y}_w)\right]^2}} \tag{2}$$

where $\overline{X}w$ is the weighted mean of $X_1,. . . ,X_p$ and $\overline{Y}w$ is the weighted mean of $Y_1,. . . ,Y_p$ with the optimal weights w^*.

5. The distance $d_{LWS}\ (C_1,\ C_2)$ between the clusters C_1 and C_2 is defined as $max\ d_{LWS}\ (X,\ Y)$ over all possible observations X coming from C_1 and all possible observations Y coming from C_2.

Ị ow we describe the whole procedure of hierarchical bottom-up robust cluster analysis. The robust cluster analysis method denoted by LWS-CA is defined in the following way.

Algorithm 2: (Robust cluster analysis LWS-CA)

1. Consider each particular observation as an individual cluster.

2. Searchfor such a pair of clusters C_1 and C_2 which have the minimal value of the robust distance $d_{LWS}\ (C_1,\ C_2)$.

3. The pair of clusters selected in step 2 is joined to a single cluster.

4. Repeat steps 2 and 3 until a suitable number of clusters is found, using the gap statistic criterion (Tibshirani et al. , 2001).

TheLWS-CAmethod corresponds to a single linkage clustering used together with a robust version of the Pearson correlation coefficient based on implicitly assigned weights. The stopping rule based on the gap statistic compares the inter-cluster variability with the intra-cluster variability in the data. Alternatively, the user may require a fixed value of the final number of resulting clusters before the computation.

4. ROBUST LOGISTIC REGRESSION

The logistic regression is a basic tool for modeling trend of a binary variable depending on one or several regressors (continuous or categorical). From the statistical point of view, it is the most commonly used special case of a generalized linear model. At the same time, it is also commonly used as a method for classification analysis (Agresti, 1990).

The maximum likelihood estimation of parameters in the logistic regression is known to be too vulnerable to the presence of outliers. Moreover, the logistic regression does not assume errors in the regressors, which may be an unrealistic assumption in many applications. Outliers appear in the logistic regression context quitecommonly as measurement errors or can emerge as typing errors both in regressors and the response (Buonaccorsi, 2010). This section proposes a novel robust estimator for parameters of logistic regression denoted based on the least weighted squares estimator (Section 2) and derive its breakdown point.

Christmann (1994), who used the least median of squares (LMS) method for estimating parameters in the logistic regression model and proved the estimator to possess the maximal possible breakdown point. Nevertheless, the low efficiency of the LMS estimator has been reported as unacceptable; see Shertzer and Prager (2002) for a management application. Čížek (2008) advocated highly robust estimation of parameters of the logistic regression, because they have the potential to have a relatively low bias without a need for a bias correction, which would be necessary for M-estimators. A nonparametric alternative to the logistic regression is known as the multifactor dimensionality reduction proposed by Ritchie et al. (2001).

Let us now consider the model of the logistic regression with a binary response $Y=(Y_1,..., Y_n)^T$. Its values equal to 1 or 0 can be interpreted as a success (or failure, respectively) of a random event. The probability of success for the i-th observation $(i=1,...,n)$ is modeled as a response of independent variables, which can be either continuous or discrete. The regressors are denoted as $X_i=(X_1,..., X_{pi})^T$ for $i=1,...,n$. The conditional distribution of Y_i assuming fixed values of the regressors is assumed to be binomial $Bi(m_i, \pi_i)$, where m_i is a known positive integer and probability π_i depends on regression parameters $\beta_1,...,\beta_p$ for $i=1,...,n$ through

$$log\frac{\pi_i}{1-\pi_i}=\beta_1 X_{1i}+...+\beta_p X_{pi}. \qquad (3)$$

We introduce the notation

$$\widetilde{X}_i = X_i\left(\frac{Y_i}{m_i}(1-Y_i)\right)^{1/2} \text{ and}$$

$$\widetilde{Y}_i = log(\pi_i/(1-\pi_i))X_i\left(\frac{Y_i}{m_i}(1-Y_i)\right)^{1/2} \qquad (4)$$

for i=1,…,n.

Definition 2. We define the least weighted logistic regression (LWLR) estimator in the model (3) in the following way. We consider the transformations (4), where the unknown probabilities π_i $(i=1,...,n)$ are estimated by the maximum likelihood method. The LWLR is defined as the LWS estimator with the adaptive quantile-based weights computed for the data

$$(\widetilde{X}_1^T,\widetilde{Y}_1)^T,...,(\widetilde{X}_n^T,\widetilde{Y}_n)^T. \qquad (5)$$

The LWLR estimator has a high breakdown point as evaluated in the following theorem, which can be proven as an analogy of the result of Christmann (1994).

Theorem 1. We assume the values m_i are assumed to be reasonably large. Further, we assume technical assumptions of Christmann (1994). Then, the breakdown point of the LWLR estimator computed with the data-dependent adaptive weights of Čížek (2011) is equal to:

$$\frac{\left\lceil \frac{n}{2} \right\rceil - p + 2}{n} \qquad (6)$$

where $\left[\frac{n}{2}\right]$ denotes the integer part of *n/2*, defined as the largest integer smaller or equal *to n/2.*

Let us discuss using the robust logistic regression as a classification method into two groups. The observed data are interpreted as a training data base with the aim to learn a classification rule. Let p_i defined by

$$p_i = e^{b_1 X_{1i}+\ldots+b_p X_{pi}} / \left(1 + e^{b_1 X_{1i}+\ldots+b_p X_{pi}}\right) \qquad (7)$$

denote the probability that the *i*-th observation belongs to the group 1, where $(b_1, \ldots, b_p)^T$ is the LWLR estimator of $(\beta_1, \ldots, \beta_p)^T$. Theoretical textbooks (e. g. Jaakkola, 2013) commonly recommend to use the logistic regression to classify a new observation to the group 1, if and only if

$$p_i > 1/2, \qquad (8)$$

which is equivalent to

$$log \frac{p_i}{1-p_i} > 0. \qquad (9)$$

However, this may be very unsuitable in some situations, especially if a large majority of the training data belongs to one of the two given groups.

A more efficient classification rule is obtained by replacing the classification rule (9) by the rule $p_i > c$, where the constantc is determined in the optimal way, that it minimizes the total classification error. We point out that it is equivalent to maximizing the Youden's index *I* (Youden, 1950) defined as:

$$I = sensitivity + specificity - 1, \qquad (9`)$$

where sensitivity defined as the probability of a correct classification of a successful observation and specificity is the probability of a correct classification of an unsuccessful observation. The optimization of the threshold c in the classification context is common for neural networks, as it will be described in Section 6. 2.

5. ROBUST MULTINOMIAL LOGISTIC CLASSIFICATION ANALYSIS

Classification analysis into several (more than two) groups is a common task in management applications. We extend the robust logistic regression estimator LWLR of Section 5 to a robust multinomial logistic classification analysis into several groups. The new method is insensitive to the presence of contamination in the data.

First, we recall the multinomial logistic regression, which is an extension of the logistic regression (3) to a model with a response with several different. We assume the total number of *n* measurements denoted as $(X_{11},\ldots,X_{1n})^T, \ldots, (X_{p1},\ldots,X_{pn})^T$. Each of them belongs to one of *K* different groups. The index variable $Y=(Y_1,\ldots,Y_n)^T$ contains the code *1,…,K* corresponding to the group membership of each measurement. We consider a model assuming that *Y* follows a multinomial distribution with *K* categories.

Definition 3. The data are assumed as *K* independent random samples of *p*-variate data. The multinomial logistic regression model is defined as

$$log \frac{p(Y_i = k)}{p(Y_i = K)} = \beta_{k1} X_{1i} + \ldots + \beta_{kp} X_{pi}, i = 1,\ldots,n,$$

$$k=1,\ldots, K-1 \qquad (10)$$

where $P(Y_i=k)$ denotes the probability that the observation Y_i belongs to the k-th group.

Here each of the groups $1,...,K$ has its own set of regression parameters, expressing the discrimination between the particular group and the reference group K. We also point out that the value

$$log\frac{p(Y_i=k)}{p(Y_i=l)}, \tag{11}$$

which compares the probability that the observation Y_i belongs to the k-th group with the probability that Y_i belongs to the l-th group $(k=1,...,K,\ l=1,.\ ,,,,K)$,can be evaluated as

$$log\frac{p(Y_i=k)}{p(Y_i=l)}=\beta_{k1}X_{li}+...+\beta_{kp}X_{pi}-\beta_{l1}X_{li}-$$

$$...-\beta_{lp}X_{pi}. \tag{12}$$

Our aim is to use the observed data as a training set to learn a robust classification rule allowing to assign a new observation $Z=(Z_1,...,Z_p)^T$ to one of the K given groups. As a solution, we define the robust multinomial logistic classification analysis based on a robust estimation computed separately in the total number of K-1 particular models (10).

Definition 4. In the model (10), let p_k denote the estimate of the probability that the observation Z belongs to the k-th group for $k=1,...,K$ by replacing the unknown regression parameters in (10) by their maximum likelihood estimates. Using the notation

$$v_{ki}=(m_ip_k(1-p_k))^{\frac{1}{2}},\widetilde{X}_{ki}=v_{ki}X_i$$

and

$$\widetilde{Y}_{ki}=v_{ki}log(p_k/(1-p_k)) \tag{13}$$

for $i=1,...,n$, the robust estimator is obtained as the LWS estimator with adaptive quantile-based weights in the model

$$(\widetilde{X}^T_{k1},\widetilde{Y}_{k1})^T,...,(\widetilde{X}^T_{kn},\widetilde{Y}_{kn})^T. \tag{14}$$

The robust multinomial logistic classification analysis (RMLCA) assigns a new observation Z to the group k, if and only if

$$p_k \geq p_j \, for\ all\ k=1,...,K,\ j\neq k. \tag{15}$$

6. NEURAL NETWORKS

6. 1. Principles of neural networks

Neural networks represent an important predictive data miningtool with a high flexibility and ability to be applied to a wide variety of complex models (Hastie et al. , 2001). They have become increasingly popular as an alternative to statistical approaches for classification (Dutt-Mazumder et al. , 2011). In management applications, neural networks are commonly used in risk management, market segmentation, classification of consuming spending patterns, sale forecasts, analysis of financial time series, quality control, or strategic planning; see the survey by Hakimpoor et al. (2011) for a full discussion. However, Krycha and Wagner (1999) warned that most papers on management applications 'do not report completelyhow they solved the problems at hand by the neural network approach'.

Neural networks are biased under the

presence of outliers and a robust estimation of their parameters is desirable (Beliakov et al. , 2011). In general, neural networks require the same validation steps as statistical methods, because they involve exactly the same sort of assumptions as statistical models (Fernandez, 2003). Actually, neural networks need an evaluation even more than the logistic regression, because more complicated models are more vulnerable to overfitting than simpler models. Unfortunately, a validation is often infeasible for complex neural networks. This leads users to a a tendency to believe that neural networks do not require any statistical assumptions or that they are model-free (Rusiecki, 2008), which may lead to a wrong feeling that they do not need any kind of validation.

We mention also some other weak points of training neural networks. They are often called black boxes, because their numerous parameters cannot be interpreted clearly and it is not possible to perform a variable selection by means of testing the statistical significance of individual parameters. Moreover, learning reliable estimates of the parameters requires a very large number of observations, especially for data with a large number of variables. Also overfitting is a common shortage of neural networks, which is a consequence of estimating the parameters entirely over the training data without validating on an independent data set.

Different kinds of neural networks can be distinguished according to the form of the output and the choice of the activation function and different tasks require to use different kinds of networks. In this work, we consider two most common kinds of multilayer perceptrons, which are also called multilayer feedforward networks. Their training is most commonly based on the back-propagation algorithm. Section 7. 2 discusses neural networks for a supervised classification task, i. e. networks with a binary output, and explains their link to the logistic regression. Section 7. 3. disusses neural networks for function approximation, i. e. networks wih a continuous output, and studies a robust approach to their fitting. A new robust estimation tool suitable for some types of such neural networks is described in Section 7. 4 as the nonlinear least weighted squares estimator.

6. 2. Neural networks for classification

We discuss supervised multilayer feedforward neural networks employed for a classification task. The input data are coming from *K* independent random samples (groups) of *p*-dimensional data and the aim is to learn a classification rule allowing to classify a new measurement into one of the groups. In this section, we also describe the connection between a classification neural network and logistic regression.

The network consists of an input and output layer of neurons and possibly of one or several hidden layers, which are mutually connected by edges. A so-called weight corresponds to each observed variable. This weight can be interpreted as a regression parameter and its value can be any real (also negative) number. In the course of the process of fitting the neural network, the weights connecting each neuron with one of neurons from the next layers are optimized. A given activation function is applied on the weighted inputs to determine the output of the neural network.

First, let us consider a neural network constructed for the task of classification to two groups with no hidden layers. The output of the network is obtained in the form

$$f(x) = g(w^T x + b) \qquad (16)$$

where $x \in R^p$ represents the input data, R denotes the set of all real numbers, w is the vector of weights, and b is a constant (intercept). Simple neural networks usually use the activation function g as a binary function or a monotone function, typically a sigmoid function such as the logistic function or hyperbolic tangent.

If the logistic activation function

$$g_1(x) = \frac{1}{1+e^{-x}}, x \in R, \qquad (17)$$

is applied, the neural networks is precisely equal to the model of logistic regression (Dreiseitl & Ohno-Machado, 2002). This special case of a neural network is different from the logistic regression only in the method for estimating the (regression) parameters. Moreover, if the hyperbolic tangent activation functionas

$$g_2(x) = tanh(x), x \in R, \qquad (18)$$

the neural network without hidden layers is again equal to the model of the logistic regression. Moreover, it can be easily proven that $g_2(x) = 2g_1(2x) + 1$ for any real x.

An influence of outliers on classification analysis performed by neural networks has been investigated only empirically (e. g. Murtaza et al. , 2010). Classification neural networks are sensitive to the presence of outliers, but a robust version has not been proposed. For the classification to two groups, the sensitivity of the neural networks is a consequence of their connection to the logistic regression. Thanks to the connection between neural networks and logistic regression, the robust logistic regression estimator LWLR (Section 5) represents at the same time a robust method for fitting classification neural networks with no hidden layers.

Moreover, references on the sensitivity analysis of neural networks have not analyzed other situations, which negatively influence training neural networks (cf. Yeung et al. , 2010). An example is multicollinearity, which is a phenomenon as harmful for neural networks as for the logistic regression.

Neural networks for classification to K groups are commonly used with one or more hidden layers. Such networks use a binary decision in each node of the last hidden layer driven by a sigmoid activation function. Therefore, they are sensitive to the presence of outliers as well as the networks for classification to two groups.

6. 3. Neural networks for function approximation

Neural networks are commonly used as a tool for approximating a continuous real function. In management, multilayer feedforward networks are often used mainly for predictions, e. g. for modeling and predicting market response (Gruca et al. , 1999) or demand forecasting (Efendigil et al. , 2009). Most commonly, they use an identical activation function, i. e. they can be described as

$$f(x) = g(w^T x + b), x \in R^p. \qquad (19)$$

An alternative approach to function approximation is to use radial basis function neural networks. Now we need to describe the back-propagation algorithm for function approximation networks, which is at the same time commonly used also for the classification networks of Section 6. 2.

The back-propagation algorithm for neural networks employed for function approximation minimizes the total error computed across all data values of the training data set. The algorithm is based on the least squares method, which is optimal for normally distributed random errors in the data (Rusiecki, 2008). After an initiation of the values of the parameters, the forward propagation is a procedure for computing weights for the neurons sequentially in particular hidden layers. This leads to computing the value of the output and consequently the sum of squared residuals computed for the whole training data set. To reduce the sum of squared residuals, the network is sequentially analyzed from the output back to the input. Particular weights for individual neurons are transformed using the optimization method of the steepest gradient. However, there are no diagnostic tools available, which would be able to detect a substantial information in the residuals, e. g. in the form of their dependence, heteroscedasticity, or systematic trend.

A robust version of fittingmultilayer feedforward networks for the task of function approximation for contaminated data was described only for specific kinds of neural networks. Rusiecki (2008) studied neural networks based on robust multivariate estimation using the minimum covariance determinant estimator. Chen and Jain (1994) investigated M-estimators and Liano (1996) studied the influence function of neural networks for function approximation as a measure of their robustness.

Jeng et al. (2011) or Beliakov et al. (2012) studied neural networks based on nonlinear regression. Then, instead of estimating the parameters of their neural networks by means of the traditional nonlinear least squares, they performed the LTS estimation. In this paper, we propose an alternative approach based on the LWS estimator. Because such approach is general and not connected to the context of neural networks, we present the methodology as a self-standing Section 6. 4.

6. 4. Nonlinear least weighted squares regression

Let us consider the nonlinear regression model

$$Y_i=f(\beta_1 X_{1i}+\ldots+\beta_p X_{pi})+e_i,\ i=1,\ldots,n, \qquad (20)$$

where $Y=(Y_1,\ldots,Y_n)^T$ is a continuous response, $(X_{11},\ldots,X_{1n})^T,\ldots,$ $(X_{p1},\ldots,X_{pn})^T$ regressors and f is a given nonlinear function. Let $u_{(i)}(b)$ denote a residual corresponding to the i-th observation for a given estimator $b=(b_1,\ldots,b_p)^T \in R^p$ of the regression parameters $(\beta_1,\ldots,\beta_p)^T$. We consider the residuals arranged in ascending order in the form

$$u^2_{(1)}(b) \le u^2_{(2)}(b) \le \ldots \le u^2_{(n)}(b) \qquad (21)$$

We define the least weighted squares estimator of the parameters in the model (20) as

$$arg\,min \sum_{i=1}^{n} w_i u^2_{(i)}(b), \qquad (22)$$

where the argument of the minimum is computed over all possible values of $b=(b_1,\ldots,b_p)^T$ and where $w_1,\ldots,w_n$ are magnitudes of weights determined by the user, e. g. linearly decreasing or logistic weights (Kalina, 2012a). Ị ow we describe the algorithm for computing the solution of (23) as an adaptation of the LTS algorithm for the linear regression (cf. Rousseeuw & van Driessen, 2006).

Algorithm 3: (Nonlinear least weighted squares estimator).

1. Set the value of a loss function to +∞. Select randomly p points, which uniquely determine the estimate b of regression parameters β.

2. Evaluate residuals for all observations. Assign the weights to all observations based on (21).

3. Compare the value of $\sum_{i=1}^{n} w_i u_{(i)}^2(b)$ computed with the resultingweights with the current value of the loss function. If $\sum_{i=1}^{n} w_i u_{(i)}^2(b)$ is larger, go to step 4. Otherwise go to step 5.

4. Set the value of the loss function to $\sum_{i=1}^{n} w_i u_{(i)}^2(b)$ and store the values of the weights. Find the nonlinear regression estimator of $\beta_1, \ldots, \beta_p$ by weighted least squares (Seber & Wild, 1989) using these weights. Go back to steps 2 and 3.

5. Repeatedly (10 000 times) perform steps 1 through 4. The output (optimal) weights are those giving the global optimum of the loss function $\sum_{i=1}^{n} w_i u_{(i)}^2(b)$ over all repetitions of steps 1 through 4.

7. CONCLUSION

This paper fills the gap of robust statistical methodology for data mining by introducing new highly robust methods based on implicit weighting. Robust statistical methods are established as an alternative approach to certain problems of statistical data analysis.

Data mining methods are routinely used by practitioners in everyday management applications. This paper persuades the readers that robustness is a crucial aspect of data mining which has remained only a little attention so far, although sensitivity of data mining methods to the presence of outlying observations has been repeatedlyreported as a serious problem (Wong, 1997; Yeung et al. , 2010). On the other hand, an experience shows that too sophisticated data mining methods are not very recommendable for practical purposes (Hand, 2006) and simple methods are usually preferable.

This paper introduces new tools for the robust data mining using an intuitively clear requirement to down-weight less reliable observations. The robust cluster analysis LWS-CA (Section 3), the robust logistic regression LWLR (Section 4), and the robust multinomial logistic classification analysis (Section 5) are examples of newly proposed methods which can be understood as part of the robust data mining framework.

The logistic regression is commonly described as a white-box model (Dreiseitl & Ohno-Machado, 2012), because it offers a simpler interpretation compared to the neural networks characterized as a black box. Because robust methods have been mostly studied for continuous data, our proposal of robust logistic regression is one of pioneering results on robust estimation in the context of discrete data.

In management applications, neural networks are quite commonly used, but still less frequently compared to the logistic regression. Also neural networks are sensitive to the presence of outlying measurements. This paper discusses the importance of robust statistical methods for training neural networks (Section 6) and proposes a robust approach to neural networksfor function approximation. Such approach still requires an extensive evaluation with the aim to detect a possible overfitting by means of a cross-validation or boostrap methodologies. Moreover, robust neural networks for classification purposes based on robust back-propagation remains to be an open problem. A classification neural network is characterized as a generalization of the logistic regression, although the two models do not use the same method for parameter estimation. Another possibility for a robust training of classification neural networks would be to propose a robust back-propagation based on the minimum weighted covariance determinant estimator proposed by Kalina (2012b).

In this paper, we do not discuss a robust information extraction from high-dimensional data, which represents a very important and complicated task. It is true that numerous data mining methods suffer from the so-called curse of dimensionality. Also neural networks have not been adapted for high-dimensional data; it can be rather recommended to use alternative approaches for a high-dimensional supervised classification (e. g. Bobrowski & Łukaszuk, 2011).

The research was supported by the grant 13-01930S of the Grant Agency of the Czech Republic

ВИСОКО РОБУСТНИ МЕТОДИ ИСТРАЖИВАЊА ПОДАТАКА

Јан Калина

Извод

Овај се рад бави високо робустним методама екстракције информација из података, уз посебну пажњу посвећену методима погодним за примену у менаџменту. Осетљивост доступних метода истражива, на присуство екстремних резултата мерења у полазној бази, је дискутована као основни недостатак разматраних метода. Овај рад предлаже неколико новијих робустних метода истраживања података, које се заснивају на идеји имплицитних тежинских коефицијената индивидуалних вредности података. Посебно се предлаже новији робустни метод хијерархијске анализе кластера, који је популаран метод анализе података и учења мрежа. Даље, предлаже се робустни метод за процену параметара током логистичке регресије. Ова идеја се шири ка робустној мултиномијалној логистичкој анализи класификације. На крају, дискутује се осетљивост неуронских мрежа на присуство шума екстремних података у полазној бази. Предлаже се метод робустног тренинга неуронских мрежа у циљу апроксимације функције, која има форму робустног алата процене у нелинеарној регресионој анализи.

Кључне речи: Истраживање података, Робустна статистика, Вишедимензиони подаци, Анализа кластера, Логистичка регресија, Неуронске мреже

References

Agresti A. (1990). Categorical data analysis. Wiley, New York.

Beliakov G. , Kelarev A. , & Yearwood J. (2012). Robust artificial neural networks and outlier detection. Technical report, arxiv. org/pdf/1110. 1069. pdf (downloaded November 28, 2012).

Bobrowski L. , & Łukaszuk T. (2011). Relaxed linear separability (RLS) approach to feature (gene) subset selection. In Xia X. (Ed.). Selected works in bioinformatics. InTech, Rijeka, 103-118.

Brandl B. , Keber C. , & Schuster M. G. (2006). An automated econometric decision support system. Forecast for foreign exchange trades. Central European Journal of Operations Research, 14 (4), 401-415.

Briner R. B. , Denyer D. , & Rousseau D. M. (2009). Evidence,based management. Concept cleanup time? Academy of Management Perspectives, 23 (4), 19-32.

Buonaccorsi J. P. (2010). Measurement error. models, methods, and applications. Boca Raton. Chapman & Hall/CRC.

Chae S. S. , Kim C. , Kim J. ,M. , & Warde W. D. (2008). Cluster analysis using different correlation coefficients. Statistical Papers, 49 (4), 715-727.

Chen D. S. , & Jain R. C. (1994). A robust back propagation learning algorithm for function approximation. IEEE Transactions on Neural Networks, 5 (3), 467-479.

Christmann A. (1994). Least median of weighted squares in logistic regression with large strata. Biometrika, 81 (2), 413-417.

Čížek P. (2011). Semiparametrically weighted robust estimation of regression models. Computational Statistics & Data Analysis, 55 (1), 774-788.

Čížek P. (2008). Robust and efficient adaptive estimation of binary,choice regression models. Journal of the American Statistical Association, 103 (482), 687-696.

Davies P. L. , & Gather U. (2005). Breakdown and groups. Annals of Statistics, 33 (3), 977-1035.

Dreiseitl S. , & Ohno,Machado L. (2002). Logistic regression and artificial neural network classification models. A methodology review. Journal of Biomedical Informatics, 35, 352-359.

Dutt,Mazumder A. , Button C. , Robins A. , & Bartlett R. (2011). Neural network modelling and dynamical systém theory. are they relevant to study the governing dynamics of association football players? Sports Medicine, 41 (12), 1003-1017.

Efendigil T. , Önüt S. , & Kahraman C. (2009). A decision support system for demand forecasting with artificial support networks and neuro,fuzzy models. A comparative analysis. Expert Systems with Applications, 36 (3), 6697-6707.

Fayyad U. , Piatetsky,Shapiro G. , & Smyth P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17 (3), 37-54.

Fernandez G. (2003). Data mining using SAS applications. Boca Raton. Chapman & Hall/CRC.

Gao J. , & Hitchcock D. B. (2010). James,Stein shrinkage to improve k,means cluster analysis. Computational Statistics & Data Analysis, 54, 2113-2127.

García,Escudero L. A. , Gordaliza A. , San Martín R. , van Aelst S. , & Zamar R. (2009). Robust linear clustering. Journal of the Royal Statistical Society, B71 (1), 301-318.

Gruca T. S. , Klemz B. R. , & Petersen E. A. F. (1999). Mining sales data using a neural network model of market reponse. ACM SIGKDD Explorations Newsletter, 1

(1), 39-43.

Gunasekaran A. , & Ị gai E. W. T. (2012). Decision support systems for logistic and supply chain management. Decision Support Systems and Electronic Commerce, 52 (4), 777-778.

Hakimpoor H. , Arshad K. A. B. , Tat H. H. , Khani N. , & Rahmandoust M. (2011). Artificial neural networks' applications in management. World Applied Sciences Journal, 14 (7), 1008-1019.

Hand D. J. (2006). Classifier technology and the illusion of progress. Statistical Science, 21 (1), 1-15.

Hastie T. , Tibshirani R. , Friedman J. (2001). The elements of statistical learning. Springer, New York.

Hekimoglu S. , Erenoglu R. C. , & Kalina J. (2009). Outlier detection by means of robust regression estimators for use in engineering science. Journal of Zhejiang University, Science A, 10 (6), 909-921.

Jaakkola T. S. (2013). Machine learning. http. //www. ai. mit. edu/courses/6. 867-f04/lectures/lecture-5-ho. pdf (downloaded January 4, 2013).

Jeng J. ,T. , Chuang C. ,T. , & Chuang C. ,C. (2011). Least trimmed squares based CPBUM neural networks. Proceedings International Conference on System Science and Engineering ICSSE 2011, IEEE Computer Society Press, Washington, 187-192.

Kalina J. (2012a). Implicitly weighted methods in robust image analysis. Journal of Mathematical Imaging and Vision, 44 (3), 449-462.

Kalina J. (2012b). On multivariate methods in robust econometrics. Prague Economic Papers, 21 (1), 69-82.

Kalina J. (2011). Some diagnostic tools in robust econometrics. Acta Universitatis Palackianae Olomucensis Facultas Rerum Naturalium Mathematica ,50 (2), 55-67.

Krycha K. A. , & Wagner U. (1999). Applications of artificial neural networks in management science. A survey. Journal of Retailing and Consumer Services, 6, 185-203.

Liang K. (2005). Clustering as a basis of hedge fund manager selection. Technical report, University of California, Berkeley, cmfutsarchive/HedgeFunds/hf_managerselection. pdf (downloaded December 20, 2012).

Liano K. (1996). Robust error measure for supervised neural network learning with outliers. IEEE Transactions on Neural Networks, 7 (1), 246-250.

Maronna R. A. , Martin R. D. , & Yohai V. J. (2006). Robust statistics. Theory and methods. Chichester. Wiley.

Martinez W. L. , Martinez A. R. , & Solka J. L. (2011). Exploratory data analysis with MATLAB. Second edition. Chapman & Hall/CRC, London.

Mura L. (2012). Possible applications of the cluster analysis in the managerial business analysis. Information Bulletin of the Czech Statistical Society, 23 (4), 27-40. (In Slovak.)

Murtaza N. , Sattar A. R. , & Mustafa T. (2010). Enhancing the software effort estimation using outlier elimination methods for agriculture in Pakistan. Pakistan Journal of Life and Social Sciences, 8 (1), 54-58.

Nisbet R. , Elder J. , Miner G. (2009). Handbook of statistical analysis and data mining applications. Elsevier, Burlington.

Punj G. , & Stewart D. W. (1983). Cluster analysis in marketing research. Review and suggestions for applications. Journal of Marketing Research, 20 (2), 134-148.

Ritchie M. D. , Hahn L. W. , Roodi N. , Bailey L. R. , Dupont W. D. , Parl F. F. , & Moore J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. American Journal of Human Genetics, 69 (1), 138-147.

Rousseeuw P. J. , & van Driessen K. (2006). Computing LTS regression for large data sets. DataMining and Knowledge Discovery,12 (1), 29-45.

Rusiecki A. (2008). Robust MCD,based backpropagation learning algorithm. In Rutkowski L, Tadeusiewicz R. , Zadeh L. , Zurada J. (Eds.). Artificial Intelligence and Soft Computing. Lecture Notes in Computer Science, 5097, 154-163.

Salibián,Barrera M. (2006). The asymptotics of MM,estimators for linear regression with fixed designs. Metrika, 63, 283-294.

Schäfer J. , & Strimmer K. (2005). A shrinkage approach to large,scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology, 4 (1), Article 32, 1-30.

Seber A. F. , & Wild C. J. (1989). Nonlinear regression. Wiley, New York.

Shertzer K. W. , & Prager M. H. (2002). Least median of squares. A suitable objective function for stock assessment models? Canadian Journal of Fisheries and Aquatic Sciences, 59, 1474-1481.

Shin S. , Yang L. , Park K. , & Choi Y. (2009). Robust data mining. An integrated approach. In Ponce J. , Karahoca A. (Eds.). Data mining and knowledge discovery in real life applications. I-Tech Education and Publishing, New York.

Soda P. , Pechenizkiy M. , Tortorella F. , & Tsymbal A. (2010). Knowledge discovery and computer,based decision support in biomedicine. Knowledge discovery and computer,based decision support in biomedicine. Artificial Intelligence in Medicine, 50 (1), 1-2.

Stigler S. M. (2010). The changing history of robustness. American Statistician, 64 (4), 277-281.

Svozil D. , Kalina J, Omelka M. , & Schneider B. (2008). DNA conformations and their sequence preferences. Nucleic Acids Research, 36 (11), 3690-3706.

Tibshirani R. , Walther G. , & Hastie T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society Series B, 63 (2), 411-423.

Vintr T. , Vintrová V. , & Řezanková H. (2012). Poisson distribution based initialization for fuzzy clustering. Neural Ị etwork World, 22 (2), 139-159.

Víšek J. Á. (2001). Regression with high breakdown point. In Antoch J. , Dohnal G. (Eds.). Proceedings of ROBUST 2000, Summer School of JČMF, JČMF and Czech Statistical Society, Prague, 324-356.

Yeung D. S. , Cloete I. , Shi D. , & Ị g W. W. Y. (2010). Sensitivity analysis for neural networks. Springer, New York.

Youden W. J. (1950). Index for rating diagnostic tests. Cancer 3, 32-35.

Zvárová, J. , Veselý A. , & Vajda I. (2009). Data, information and knowledge. In P. Berka, J. Rauch and D. Zighed (Eds.), Data mining and medical knowledge management. Cases and applications standards. IGI Global, Hershey, 1-36.