www.sjm06.com

Serbian Journal of Management 9 (1) (2014) 131 - 144

Serbian Journal of Management

ON ROBUST INFORMATION EXTRACTION FROM HIGH-DIMENSIONAL DATA

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 17 February 2014; accepted 30 March 2014)

Abstract

Information extraction from high-dimensional data represents an important problem in current applications in management or econometrics. An important problem from a practical point of view is the sensitivity of machine learning methods with respect to the presence of outlying data values, while numerical stability represents another important aspect of data mining from high-dimensional data. This paper gives an overview of various types of data mining, discusses their suitability for high-dimensional data and critically discusses their properties from the robustness point of view, while we explain that the robustness itself is perceived differently in different contexts.Moreover, we investigate properties of a robust nonlinear regression estimator of Kalina (2013).

Keywords: Data mining, high-dimensional data, robust econometrics, outliers, machine learning

1. DATA MINING FROM HIGH-DIMENSIONAL DATA

High-dimensional data are often encountered in management applications with the aim to perform a decision making, which can be described as selecting an activity or series of activities among several alternatives (Martinez et al., 2011). Data mining methods for information extraction from high-dimensional data represent an important tool allowing to find answers to given questions concerning a fixed database or to generate hypotheses from a random sample.

High-dimensional data are usually understood to have a form of a data set with a large number of observations and/or a large number of variables. Statisticians usually consider a situation with a small number of

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

DOI:10.5937/sjm9-5520

observations (Hall et al., 2005), while the term big data is used in computer science in a broader sense for such data, if there is an additional requirement to automate the analysis within e.g. online applications. Indeed, information extraction from data with a large number of variables is complicated even in situations with a large number of observations.

An important area of applications of high-dimensional information extraction consists in 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 (Kalina et al., 2013). Such partially or fully automatic systems are capable to solve a variety of complex tasks, to analyze a large database containing different information components, to extract information of different types, and deduce conclusions from them in management or econometric applications (Brandl et al., 2006).Nevertheless, the largest applications of high-dimensional information extraction can be found in molecular genetics or image analysis.

Standard multivariate statistical methods turn out to be unreliable for high-dimensional data. An intensive current research in statistics has the aim to propose new multivariate methods tailor-made for classification of high-dimensional data, if the number of variables exceeds the number of observations. Several works have shown that an analysis starting with a dimensionality reduction is suboptimal, although it remains to be the most common approach (Greenland, 2000). There is an urgent demand for new reliable methods for high-dimensional data in econometric and management applications.

A high dimension of the data is a major problem also in data mining applications. A management database, e.g. a customer analytical record (CAR), may contain a huge number of variables reported for a large number of units, while the database of units may correspond to the entire population. Therefore, data mining requires tailor made methods suitable for the analysis of high-dimensional data, while multivariate statistics is traditionally focused only on data with a small dimension. We can say that a high-dimensional data set does not even need a (purely) statistical analysis and data mining is more suitable for information extraction from high-dimensional data than classical statistical methods.

At any case, specific methods for data mining from high-dimensional data are only at the beginning of their development and there is no unanimity concerning the suitability of particular methods in different situations (Kalina, 2014). Thus, the situation seems rather chaotic and no systematic comparison of the performance of particular methods in different applications has been presented (Turchi et al., 2013). It is also possible to criticize available software for lack of reliability or delay in the implementation of newly proposed specific methods for the information extraction from high-dimensional data.

This paper has the following structure. Section 2 discusses various definitions of robustness.Section 3 gives an overview of robust methods for dimensionality reduction. While we described robustness aspects of multilayer perceptrons in Kalina (2013), other types of neural networks are discussed in Section 4 and Section 5 is devoted to support vector machines. We contribute to the research direction of robust data mining in Section 6, which investigates properties of

the robust nonlinear regression estimator from Kalina (2013). Throughout the paper, examples of applications in management or econometrics are given.

2. THE PROBLEM OF ROBUSTNESS

The concept of robustness has been understood in different ways in robust statistics, computer science, numerical mathematics, or optimization. In a broader definition, robustness is insensitivity to violations of assumptions or to deviation from a standard situation. Thus, we can perceive robustness as numerical stability or as insensitivity to the presence of noise, outlying measurements, normal distribution of the data, and high dimensionality.

Still the existing multivariate statistical methodology suitable for highly dimensional data is too sensitive (non-robust) to the presence of outlying or incorrectly measured values (Martinez et al., 2011). Robustness properties of current high-dimensional methods have been investigated e.g. by Guo et al. (2007), although these general methods have been investigated primarily in molecular genetic applications.

Robust statistics defines robustness as insensitivity to the presence of outlying measurements (outliers), which are capable to influence classical statistical methods heavily. Statisticians and econometricians have developed the robust statistical 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 (Jurečková & Picek, 2006; Kalina, 2012). Nevertheless, the majority of robust statistical methods is computationally infeasible for high-dimensional data.

In numerical mathematics, robustness can be interpreted as insensitivity to the rounding error or to small changes of the data. Let us motivate this approach to robustness by the task of solving a linear set of equation

$$Ax = b \tag{1}$$

by the least squares method. A requirement to reduce the influence of noise on the computed solution leads to a modification of the least squares method, most commonly by the Tikhonov regularization. Then, the solution is obtained as the solution of the minimization

$$\min\{||b - Ax||^2 + ||\lambda x||^2\} \tag{2}$$

over x. The corresponding set of normal equations can be formulated as

$$(A^TA + \lambda^2 I)x = Ab \tag{3}$$

and therefore the solution x has the form

$$\hat{x} = (A^T A + \lambda^2 I)^{-1} A^T b, \tag{4}$$

where I is a unit matrix. The solution is known as the ridge regression estimator (Hastie et al., 2009). The concept of robust data mining was introduced as a methodology based on robust optimization, i.e. “optimization to provide stable solutions that can be used in case of input modification” (Xanthopoulos et al., 2013).

In general, (4) can be described as a regularized version of the least squares estimator. Regularization allows to solve ill-posed or insoluble high-dimensional problems by means of additional information, assumptions, or penalization. An intensive current research in statistics has the aim to propose regularized multivariate

methods tailor-made for classification of complex data. Regularization is also the basis of support vector machines, as this will demonstrated in Section 4. General relationship between regularization and robust approaches was investigated by Jurečková and Sen (2006). Nevertheless, a regularized method is not necessarily robust. To give an illustration, Jurczyk (2012) explained that the ridge regression estimator (4) is not robust from the statistical point of view.

Combining both the numerical and statistical point of view, it is desirable for practical methods to be double robust. This concept will be presented in the context of cluster analysis. This is a different concept from the double robustness of Funk et al. (2011), which combines robustness for two different epidemiological models. The necessity of robustifying existing methods for high-dimensional applications is well known (Hubert et al., 2008). In multivariate statistics, the Mahalanobis distance can be criticized for being sensitive both to outlying measurements and to a high dimensionality.

Besides the non-robustness, we can mention several other complications, which are relevant for the information extraction from biomedical data. Other problems not covered by this paper are related to measuring instrumental variables instead of the original ones, unrealistically strong assumptions of statistical approaches, or dichotomization of continuous data (Harrell, 2001).

3. DIMENSIONALITY REDUCTION

Dimensionality reduction methods suitable for high-dimensional data include both linear and nonlinear methods. Belloni and Chernozhukov (2011) gave an overview of the methodology suitable for econometric applications. Linear methods, e.g. principal component analysis or factor analysis, are commonly based on matrix eigendecomposition. Numerically stable algorithms are available (McFerrin, 2013),but there exist such implementations in software, which fail for data with the number of variables exceeding the number of observations. Further, variable selection by means of hypothesis testing (Smyth, 2005) is a common approach. However, its primary aim in this context is to rank the variables in the order of evidence against the null hypothesis rather than to assign p-values to variables. Other approaches to dimensionality reduction include approaches based on the information theory (Furlanello et al., 2003) or variable selection simultaneously with statistical modeling, e.g. lasso (Hersterberg et al., 2008). Statisticians have a tendency to search for parsimonious models, i.e. simple models with a small set of relevant variables, which was criticized by Harrell (2001) as unjustifiable in some cases.

If the high-dimensional data are observed in two or more different groups, then it is important to know that common dimensionality reduction methods are not suitable, i.e. they are tailor-made for classification purposes. Naïve approaches to classification of high-dimensional data start by dimension reduction and proceed with a consequent classification analysis. Several comparisons of various dimension reduction techniques for the classification context were compared e.g. by Dai et al. (2006) or Suzuki and Sugiyama (2010). Zuber and Strimmer (2011) proposed a variable selection procedure for a high-dimensional regression, which takes correlation among regressors into account. The method encourages

grouping of correlated regressors and down-weights antagonistic variables.

Robust dimensionality reduction procedures include the method of Vanden Branden and Hubert (2005) called robust soft independent modelling of class analogies (RSIMCA). It is a dimension reduction technique tailor-made for the classification task. The method applies a robust principal component analysis (ROBPCA) separately on each group of the data. Here, each group is reduced to a different dimension. A new observation is classified by means of its deviations to the different robust principal component analysis (PCA) models, exploiting a robust Mahalanobis distance. Other important approaches to dimension reduction for high-dimensional data include the sliced inverse regression (Duan & Li, 1991) or minimum redundancy maximum relevance (Liu et al., 2005).

4. NEURAL NETWORKS

Machine learning methodology represents a variety of very flexible popular tools for solving various types of problems. According to the type of learning, it is commonly distinguished between supervised and unsupervised machine learning methods (Hastie et al., 2009). While multilayer perceptrons were critically reviewed in Kalina (2013), in this paper we discuss other types of neural networks together with their performance for high-dimensional data. Section 4.1 recalls briefly multilayer perceptrons, Section 4.2 is devoted to radial basis function networks, and Section 4.3 to self-organizing maps. Table 1 gives a list of software tools available for the computation of described methods within the R software package.

4.1. Multilayer Perceptron

First, we would like to disprove a common belief that multilayer perceptrons do not demand any assumptions about the probability distribution of the data. However, they do have such assumptions on the data distribution which are analogous to assumptions of statistical models. Actually, some simple special cases of neural networks are equivalent to commonly used statistical methods. Therefore, it would be important to check the assumptions, as it is common to validate the assumptions of common statistical methods. In contrary to statistical modeling, a practical data mining inclines to ignoring the assumptions (Fernandez, 2003) and the consequences of their violation. Moreover, neural networks are not even accompanied by such diagnostic tools for validating the assumptions.

Recent references described the sensitivity of neural networks with respect to the presence of outlying data points (outliers) in the data (Murtaza et al., 2010). Estimates of parameters turn out to be biased and under the presence of outliers and it is actually desirable to estimate the parameters in a different robust way in such a situation (Rusiecki, 2008). Other works studied neural

Table 1. Overview of various types of machine learning methods

Method	Type of learning	Package in software R
Multilayer perceptron	Supervised	nnet, neuralnet
Radial basis function network	Supervised	RSNNS
Self-organizing map	Unsupervised	kohonen
Support vector machine	Supervised	e1071

networks based on robust estimators of parameters in nonlinear regression (Jeng et al., 2011). The problem of robustness of multilayer percpetrons is connected also to the generalization ability of the networks, which may be improved by pruning or selecting relevant variables for the optimal learning. Fortunately, a variety of tools for both pruning and variable selection for neural networks is available (Šebesta & Tučková, 2005). In practical applications, multilayer perceptrons have been observed to be suitable also in the high-dimensional setting (Rowley et al., 1998; Zimmermann et al., 2001).

4.2. Radial Basis Function Network

A radial basis function network is able to model a continuous nonlinear function. In contrary to multilayer perceptrons, the input layer transmits a measure of distance of the data from a given point to the following layer. Such measure is called a radial function. Typically, only one hidden layer is used and an analogy of the back-propagation is used to find the optimal values of parameters. The output of the network has the form

$$f(x) = \sum_{i=1}^{n} \left\{ w_i \, exp(-\beta \| x - c_i \|^2) \right\} =$$

$$= \sum_{i=1}^{n} w_i \, exp\left\{ -\beta (x - c_i)^T (x - c_i) \right\} \qquad (5)$$

for $x \in \mathrm{R}^p$, where n is the total number of neurons in the network and c_i is a given point corresponding to the i-th neuron.

The radial basis function itself is defined as

$$\varphi(x, c_i) = \exp\{-\beta \| x - c_i \|^2\}, x \in \mathrm{R}^p, \qquad (6)$$

and the points c_i can be interpreted as centers, from which the Euclidean distances are computed.

The output (5) is a sum of weighted probability densities of the normal distribution. The training of the networks requires to determine the number of radial units and their centers and variances. The formula (5) does not contain a normalizing constant for the density of the multivariate normal distribution, but it is contained in the weights for individual neurons. The rate of converge of radial basis function networks in approximating smooth functions has been investigated e.g. in Kainen et al. (2009). Nevertheless, this type of network is less suitable for high-dimensional data (Nisbet et al., 2009).

4.3. Self-Organizing Map

Self-organizing map is a type of a neural network searching for a mapping of multidimensional data to a low-dimensional grid with a clear graphical interpretation (Kohonen, 1982). It transforms complicated nonlinear associations to geometrically simpler ones, most commonly to dimension 2. The network has the ability to organize the data and serves as an unsupervised tool for exploration and visualization of high-dimensional data and revealing associations among variables.

The network has only an input layer and an output layer or radial units with neurons geometrically arranged to a two-dimensional grid with a given topological structure, e.g. square or hexagonal. Each neuron of the input layer is connected with all neurons of the output layer. The process of learning proceeds iteratively in the following way. For a given observation, such neuron is searched for, which has the best

correspondence to the observation, i.e. which places the observation to the map so that the topology of the observed data is preserved as well as possible. This learning corresponds to a competition among neurons driven by the rule that the winner takes all, i.e. the neuron with the best reaction on the stimulus is found. The winning neurons are arranged and constitute the set of coordinates in the grid.

The final visualization depicts all observations in the grid. Such observations, which are close to each other in the original high-dimensional space, are close also in this grid. Therefore, we can say that this neural networks creates a topological map of the input variables. Thus, the method is close to multidimensional scaling. Besides, a self-organizing map may lead to revealing clusters in the data. Therefore, it may be used as a clustering procedure prior to a consequent classification analysis. There is a good experience with the stability of self-organizing maps for high-dimensional data and they are even recommended as a reasonable alternative of cluster analysis for high-dimensional data (Penn, 2005).

5. SUPPORT VECTOR MACHINES

Neural networks have been criticized for their extreme simplicity from the theoretical point of view, e.g. by Minsky already around 1968. Although neural networks are successful in practical tasks, it is commonly explained by their combination with a sophisticated heuristics. Vapnik (1995) not only explained the suboptimality of neural networks e.g. in classification tasks, but also brought a constructive alternative called support vector machine (SVM).

A SVM explicitly formalizes the concepts solved implicitly by neural networks, but a neural network does not represent a special caes of the SVM. Instead, a SVM can be considered a close relative of neural networks and an alternative approach to their training. The difference is e.g. in searching for the optimal values of the parameters, which allow the optimal prediction. Compared to heuristically based neural networks, the SVM stands on a profound mathematical background and yields considerably better results (Nisbet et al., 2009). The SVM as a supervised learning method spread quickly to various classification and regression applications and a practical interest for neural networks started to decline.

A linear SVM classifier for classification into two groups is based on searching for such linear structure (hyperplane), which maximizes the margin between the two groups. It is based on support vectors, which are defined as selected observations near the margin between the groups. The classification rule depends on the value of a parameter λ, which is responsible for the width of the margin between the groups and the smoothness of the nonlinear boundary, which separates both groups. A narrow margin corresponds to a wiggly boundary curve, which reproduces the support vectors to a large extent. On the other hand, a wide margin corresponds to a smooth boundary between both groups. It has a worse ability to classify data from the training set, but is usually better in classifying new independent observations. A suitable value of λ is determined by a cross validation.

A nonlinear SVM starts by projecting the data to a space with a larger dimension. The linear classification problem is solved there and linear boundaries in the larger space correspond to nonlinear classification

boundaries in the original space. The transformation between both spaces is granted by a kernel transformation with a positive semidefinite kernel. Searching for the optimal linear rule with the widest margin requires intensive computations of inner products in a high-dimensional space. Thanks to the so-called kernel trick, the computation is not needed to be computed explicitly for the high dimension, but it is sufficient to perform a much simpler computation of the value of the kernel applied on the original data. As an illustration, let us consider a classification into two groups with selected support vectors $x_1, \ldots, x_S$. Let their response is equal to +1 for values in group 1 and -1 for values in group 2. Then, the output of the classifier is copmuted as

$$f(x) = \sum_{i=1}^{S} w_i y_i K(x_i, x) + b, \qquad (7)$$

where K is the kernel function, $w_1, \ldots, w_S$ weights and b an intercept. The most common choice of the kernel function is the radial basis function (6).

Thanks to controling the complexity of the solution, the SVM does not suffer from the curse of dimensionality. The optimization of parameters of a SVM is based on searching for an equilibrium between a prediction ability of the model and complexity of the solution, which is expressed by means of the Vapnik-Chervonenkis dimension (VC dimension). This principle called structural risk minimization (SRM) corresponds to a regularized version of a classical statistical approach minimizing the empirical risk. At the same time, it is a correction for a finite number of observations in a certain sense. However, optimizing the values of the parameters requires a large number of observations to be available.

Concerning robust properties of the SVM, it is robust in the sense of the robust statistics based on the concept of influence function (Christmann & Van Messem, 2008). An important research topic in the last 10 years is focused on assumptions, which ensure the SVM to be consistent. It is known that the SVM is consistent under the assumption that the loss function has a specific form. It requires complicated considerations in functional spaces to derive the consistency.

Some references claim that the SVM leads to results comparable to those obtained by a much simpler model (Blankertz et al., 2008) such as regularized linear discriminant analysis (Guo et al., 2007) or linear regression. From the statistical point of view, the SVM is based on a rather complicated model. Still, it allows to obtain reliable results in high-dimensional applications, e.g. in image analysis. Vapnik (1995) applied the SVM to a task of recognizing hand-written digits in images. Later, Osuna et al. (1997) used the SVM for the face detection in gray-scale images. In a training data set containig 50 000 faces and non-faces, the method selected 2500 support vectors, which have the form of faces with the largest similarity to non-faces as well as non-faces with the largest similarity to faces. The classification rule is based only on these images completely ignoring the remaining ones. These support vectors can be considered prototypes of objects on the boundary between the group of faces and non-faces.

Bobrowski and Łukaszuk (2011) proposed an alternative method to the SVM, which relaxes the linear separability. It is suitable for high-dimensional genetic data, because the sparsity of the data in the high-dimensional space usually allows the data to

be separated linearly (by a hyperplane). The method successively removes selected variables from the model so that a good linear separation among the groups is retained. Further, the authors extended the method to censored clinical data about patient survival (Bobrowski & Łukaszuk, 2012).

6. NONLINEAR REGRESSION

The nonlinear least weighted squares (NLWS) regression estimator and an efficient algorithm for its computation were proposed in Kalina (2013). Assuming a nonlinear regression model, the estimator is based on down-weighting less reliable observations, which are found during the computation of the estimator. Now, we show two examples illustrating the potential of the method.

Example 1. We illustrate the performance of the nonlinear least weighted squares estimator on a numerical example. The data set consists of 8 data points shown in Figure 1. The nonlinear regression model is used in the form

$$Y_i = a + b\,(X_i - c)^2 + e_i, \quad i = 1, \ldots, n, \qquad (8)$$

where $Y_1, \ldots, Y_n$ are values of the response, $X_1, \ldots, X_n$ values of the regressor, a, b, and c are regression parameters and $e_1, \ldots, e_n$ are random errors.

Figure 1 shows fitted values corresponding to the least squares fit and also the least weighted squares fit with the linearly decreasing weights. The least squares fit has the tendency to fit well also influential data points. The robust fit is able to find such subset of the data points, for which there is a very good regression fit. At the same time, it down-weights data points corresponding to larger values of the regressor.

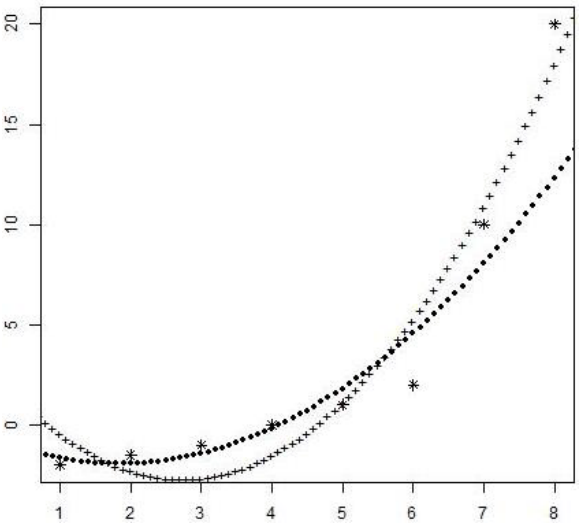

Figure 1. Nonlinear least squares (plus signs) and nonlinear least weighted squares (bullets) estimators in Example 1

Table 2 gives an evidence in favor of the algorithm for computing the NLWS estimator.

Table 2.Values of various loss functions for the least squares and nonlinear least weighted squares estimators in Example 1

Method	$\sum_{i=1}^{n} u_{(i)}^2(b)$	$\sum_{i=1}^{n} w_i u_{(i)}^2(b)$
Least squares	23.44	1.21
NLWS	70.94	0.67

The least squares estimator minimizes the value of $\sum_{i=1}^{n} u_{(i)}^2(b)$. Therefore, it may be expected that it also has a quite small value

of $\sum_{i=1}^{n} w_i u_{(i)}^2(b)$, which is the loss function of the NLWS estimator. In this example, the NLWS estimator has a much larger value of $\sum_{i=1}^{n} u_{(i)}^2(b)$ than the least squares fit. However, the algorithm used for computing the NLWS has found even a much smaller value of $\sum_{i=1}^{n} w_i u_{(i)}^2(b)$ than the least squares. This allows us to conclude that the NLWS algorithm turns out to give a reliable result.

Example 2. The purpose of this example is to illustrate the behavior of various nonlinear regression estimators for heteroscedastic data, which are shown in Figure 2. At the same time, the example reveals an undesirable property of the nonlinear least trimmed squares estimator (NLTS), which is a highly robust estimator in the nonlinear regression and extension of the least trimmed squares (LTS) estimator (Rousseeuw & van Driessen, 2006).

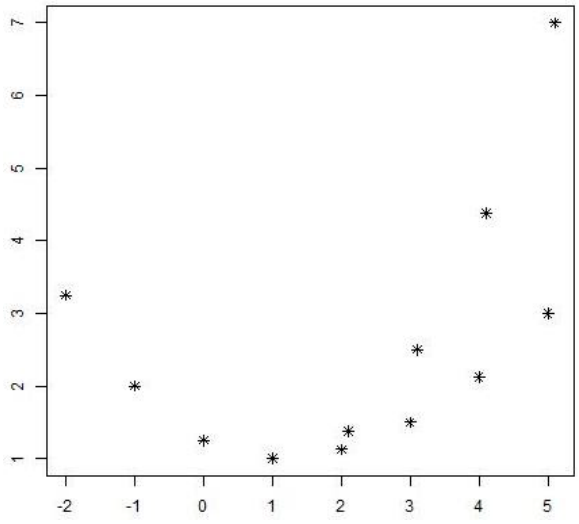

Figure 2. Original data in Example 2

We use the same model (9) as in Example 1. Figure 3 shows the results for the least squares, the NLTS (trimming away 25 % of the data points) and NLWS with linearly decreasing weights. The NLTS estimator completely ignores the heteroscedastic nature of the data and finds an unsuitable subset of the data, for which the regression fit seems very good.Such inappropriate behavior of the NLTS estimator has not been reported, but corresponds to an analogous problem of the LTS estimator in the linear regression model. The problem is associated with the high local sensitivity of the LTS estimator, which was described by Víšek (2000).

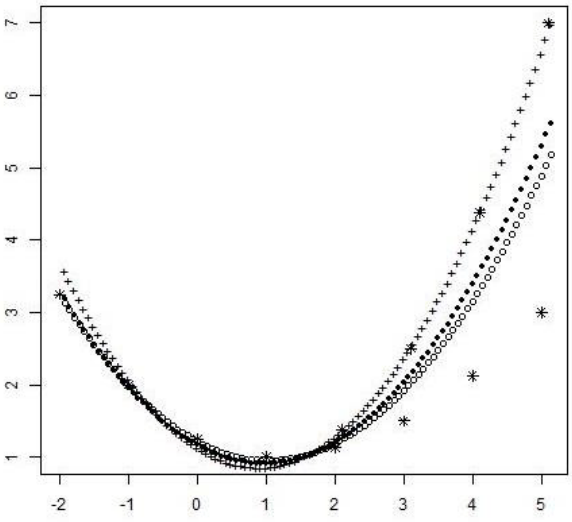

Figure 3. Various nonlinear regression estimators under heteroscedasticity in Example 2: least squares (empty circles), least trimmed squares (plus signs), and least weighted squares (full circles)

The least squares as well the NLWS estimators seem to find a more adequate regression fit also for data points with the regressor exceeding the value 2; their residuals are namely much closer to symmetry around 0. Thus, Example 2 brings an arguments in favor of the NLWS

estimator compared to the existing NLTS estimator.

To summarize, this paper recalls principles of machine learning and gives an overview of important types of methods, including multilayer perceptrons, radial basis function networks, self-organizing maps, and support vector machines. All of these methods are commonly to used to solve a variety of tasks in business and econometric applications. The paper discusses the assumptions and limitations of the methods. It follows that a robust estimation of parameters in machine learning methods is highly desirable. Furthermore, we focus on the task of function approximation by multilayer perceptrons and give an overview of existing works based on robust estimation in nonlinear regression. As an original result, we propose the NLWS estimator, describe an approximative algorithm for its computation, and show its performance on numerical examples. While the estimator is constructed to be resistant to the presence of outlying measurements in the data, there seems an advantage in assigning smaller weights to outliers compared to their complete trimming as performed by the existing NLTS estimator.

Acknowledgements

The work was supported by the Czech Science Foundation project No. 13-01930S (Robust methods for nonstandard situations, their diagnostics and implementations).

О ЕКСТРАКЦИЈИ РОБУСТНИХ ПОДАТАКА НА ОСНОВУ ВИСОКО ДИМЕНЗИОНИХ ПОДАТАКА

Jan Kalina

Извод

Екстракција информација из високо димензионих података представља веома значајан проблем у савременом примењеном менаџменту и економетрији. Значајан аспект, са практичног гледишта, је осетљивост метода учења машина у присуству екстрамних вредности података, док други аспект представља нумеричка стабилност приликом добијања информације из вишедиминзионих података.

Овај рад даје преглед типова “data mining-a”, дискутује њихову погодност за вишедимензионе податке и критички дискутује њихове особине са погледа робустности, док се сама робустност објашњава као различита перцепција у различитим концептима. Такође, врши се анализа особина робустностног нелинеарног регресионог естиматора Калина (2013).

Кључне речи: “Data mining”, високо димензиони подаци, робустна економетрија, екстреми, учење машина

References

Belloni, A., Chernozhukov, V., & Hansen, C. (2011). Inference for high-dimensional sparse econometric models. Centre for Microdata Methods and Practice working paper 41/11. [Online] Available: http://arxiv.org/pdf/1201.0220.pdf (February 12, 2014)

Blankertz, B., Tangermann, M., Popescu, F., Krauledat, M., Fazli, S., Dónaczy, M., Curio, G., & Müller, K.R. (2008). The Berlin brain-computer interface. Lecture Notes in Computer Science, 5050, 79-101.

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

Bobrowski, L.,& Łukaszuk, T. (2012). Prognostic modeling with high dimensional and censored data. Lecture Notes in Computer Science, 7377, 178-193.

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

Christmann, A., & Van Messem, A. (2008). Bouligand derivatives and robustness of support vector machines for regression. Journal of Machine Learning Research, 9, 915-936.

Dai, J.J., Lieu, L., & Rocke, D. (2006). Dimension reduction for classification with gene expression microarray data. Statistical Applications in Genetics and Molecular Biology, 5 (1), Article 6.

Duan, N., & Li, K.C. (1991). Slicing regression: A link-free regression method. Annals of Statistics, 19, 505-530.

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

Funk, M.J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M.A., & Davidian, M. (2011). Doubly robust estimation of causal effects. American Journal of Epidemiology, 173 (7), 761-767.

Furlanello, C., Serafini, M., Merler, S., & Jurman, G. (2003). Entropy-based gene ranking without selection bias for the predictive classification of microarray data. BMC Bioinformatics, 4, Article 54.

Greenland, S. (2000). When should epidemiologic regressions use random coefficients? Biometrics, 56, 915-921.

Guo, Y., Hastie, T., & Tibshirani, R. (2007). Regularized discriminant analysis and its application in microarrays. Biostatistics, 8 (1), 86-100.

Hall, P., Marron, J.S., & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society B, 67 (3), 427-444.

Harrell, F.E. (2001). Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. New York: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. Data mining, inference, and prediction. New York: Springer.

Hersterberg, T., Choi, N.H., Meier, L., & Fraley, C. (2008). Least angle and l1 penalized regression: A review. Statistics Surveys, 2, 61-93.

Hubert, M., Rousseeuw, P.J., & Van Aelst, S. (2008). High-breakdown robust multivariate methods. Statistical Science, 23, 92-119.

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, Washington: IEEE Computer Society Press, 187-192.

Jurczyk, T. (2012). Outlier detection under multicollinearity. Journal of Statistical Computation and Simulation, 82 (2), 261-278.

Jurečková, J., & Picek, J. (2006). Robust statistical methods with R. Boca Raton: Chapman & Hall/CRC.

Jurečková, J., & Sen, P.K. (2006). Robust multivariate location estimation, admissibility, and shrinkage phenomenon. Statistics & Decisions, 24, 273-290.

Kainen, P.C., Kůrková, V., & Sanguineti M. (2009). Complexity of Gaussian-radial-basis networks approximating smooth functions. Journal of Complexity, 25, 63-74.

Kalina, J. (2014). Classification analysis methods for high-dimensional genetic data. Biocybernetics and Biomedical Engineering, 34 (1), 10-18.

Kalina, J. (2013). Highly robust methods in data mining. Serbian Journal of Management, 8 (1), 9-24.

Kalina, J., Seidl, L., Zvára, K., Grünfeldová, H., Slovák, D., & Zvárová, J. (2013). Selecting relevant information for medical decision support with application to cardiology. European Journal for Biomedical Informatics, 9 (1), 2-6.

Kalina, J. (2012). On multivariate methods in robust econometrics. Prague Econ. Pap., 21, 69-82.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, 59-69.

Liu, X., Krishnan, A., & Modry, A. (2005). An entropy-based gene selection method for cancer classification using microarray data. BMC Bioinformatics, 6, Article 76.

Martinez, W.L., Martinez, A.R., & Solka, J.L. (2011). Exploratory data analysis with MATLAB. (2nd ed.). London: Chapman & Hall/CRC.

McFerrin, L. (2013). Package HDMD. [Online] Available: http://cran.r-project.org/web/packages/HDMD/HDMD.pdf (June 14, 2013)

Mosteller, F., & Tukey, J.W. (1968). Data analysis, including statistics. In G. Lindzey, E. Aronson (Eds.), Handbook of Social Psychology, Vol. 2 (pp. 80-203). New York: Addison-Wesley.

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

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, 54-58.

Osuna, E., Freund, R., & Girosi, F. (1997). Training support vector machines: An application to face detection. Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 1997, Los Alamitos: IEEE Computer Society Press, 130-136.

Penn, B.S. (2005). Using self-organizing maps to visualize high-dimensional data. Computers & Geosciences, 31 (5), 531-544.

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

Rowley, H., Baluja, S., & Kanade, S. (1998). Neural network-based face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20, 23-38.

Rusiecki, A. (2008). Robust MCD-based backpropagation learning algorithm. Lecture Notes in Computer Science, 5097, 154-163.

Šebesta, V., & Tučková, J. (2005). The extraction of markers for the training of neural network dedicated for the speech

prosody control. In S. Lecoeuche, D. Tsaptsinos (Eds.), Novel Applications of Neural Networks in Engineering International Conference on Engineering Applications of Neural Networks EANN'05, 245-250.

Smyth, G.K. (2005). Limma: linear models for microarray data. In Gentleman R., Carey V., Dudoit S., Irizarry R., Huber W. (Eds.): Bioinformatics and computational biology solutions using R and Bioconductor. Springer, New York, 397-420.

Suzuki, T., & Sugiyama, M. (2010). Sufficient dimension reduction via squared-loss mutual information estimation. Neural Computation, 25, 725-758.

Turchi, M., Perrotta, D., Riani, M., & Cerioli, A. (2013). Robustness issues in text mining. Advances in Intelligent Systems and Computing, 190, 263-272.

Vanden Branden, K., & Hubert, M. (2005). Robust classification in high dimensions based on the SIMCA method. Chemometrics and Intelligent Laboratory Systems, 79, 10-21.

Vapnik, V.N. (1995). The nature of statistical learning theory. New York: Springer.

Víšek, J.Á. (2000). On the diversity of estimates. Computational Statistics & Data Analysis, 34, 67-89.

Xanthopoulos, P., Pardalos, P.M., Trafalis, T.B. (2013). Robust data mining. Springer, New York.

Zimmermann, H.-G., Grothmann, R., & Neuneier, R. (2001). Multi-agent FX-market modeling by neural networks. Operations Research Proceedings, 2001, 413-420.

Zuber, V., & Strimmer, K. (2011). High-dimensional regression and variable selection using CAR scores. Statistical Applications in Genetics and Molecular Biology, 10 (1), Article 34.