INFORMATICA, 2011, Vol. 22, No. 1, 1–10

Large-Scale Data Analysis Using Heuristic Methods

Gintautas DZEMYDA, Leonidas SAKALAUSKAS
Vilnius University Institute of Mathematics and Informatics
Akademijos 4, LT-08663 Vilnius, Lithuania
e-mail: gintautas.dzemyda@mii.vu.lt

Received: September 2010; accepted: February 2011

Abstract. Estimation and modelling problems as they arise in many data analysis areas often turn out to be unstable and/or intractable by standard numerical methods. Such problems frequently occur in fitting of large data sets to a certain model and in predictive learning. Heuristics are general recommendations based on practical statistical evidence, in contrast to a fixed set of rules that cannot vary, although guarantee to give the correct answer. Although the use of these methods became more standard in several fields of sciences, their use for estimation and modelling in statistics appears to be still limited. This paper surveys a set of problem-solving strategies, guided by heuristic information, that are expected to be used more frequently. The use of recent advances in different fields of large-scale data analysis is promoted focusing on applications in medicine, biology and technology.

Keywords: heuristics, robust statistics, Markov model, regression, clustering, visualization.

1. Introduction

Estimation and modelling problems as they arise in many data analysis areas often turn out to be unstable and/or intractable by standard numerical methods. Such problems occur especially frequently in fitting of large data sets to a certain model according to the fit goodness function and in predictive learning. Although simplicity is desired, to be a representative of the perceived situation the resulting model may be of such a form as to make it difficult sometimes if not impossible to find its optimal solution. Given that "... all models are wrong, but some are useful " (Box, 1979), one way consists in simplifying models and procedures. However, the solutions to such simplified problems might not be satisfying. The next approach consists in looking for strategies by using readily accessible, though loosely applicable, information to control solving of the problem and involving, in an intelligent way, the trial and error experience. Such an approach had an intensive development in the heuristic framework (Tversky and Kahneman, 1974; Pearl, 1983; Michalewicz and Fogel, 2004; Sakalauskas and Zavadskas, 2009; etc.), engaging in search through a variety of methods and processes aimed at discovery. Heuristics are general recommendations based on practical evidence, in contrast to a fixed set of rules (algorithmic) that cannot vary, although guarantee to give the correct answer (Winker and

Gilli, 2004). Although the use of these methods became more standard in several fields of sciences, their use for estimation and modelling in statistics appears to be still limited. This paper surveys a set of problem-solving strategies, guided by heuristic information, that are expected to be used in large-scale data analysis more frequently.

2. Heuristics in Robust Statistics

A drawback of robust statistical techniques is the increased computational effort often needed as compared to nonrobust methods (Chiyoshi *et al.*, 2000, Meiri and Zahavi, 2006; Sakalauskas and Krarup, 2006; Nunkesser and Morell, 2010; Benati, 2011). Particularly, robust estimators possessing the exact fit property are NP-hard to compute. It means that, under the widely believed assumption that the computational complexity classes NP and P are not equal, there is no hope to compute exact solutions for large high-dimensional data sets. To tackle this problem, search heuristics are used to compute NP-hard estimators in high dimensions. Since in the studies of large data sets where influential observations or outliers maybe present, regression models based on the maximum likelihood criterion are likely to be unstable, the use of the another, minimum density power divergence criterion has been explored as a practical tool for parametric regression model building. Thus, a procedure relying on the index of similarity between the estimated regression models and on a Monte Carlo significance test of hypothesis that allows us to check the existence of outliers in the data and therefore to choose the best tuning constant for the minimum density power divergence has been developed and studied by Durio and Isaia (2011). The procedure proposed allows simultaneously to detect the presence of outliers in the data and to select the best tuning constant for the estimators. The core of the procedure relies on the concept of similarity between the estimated regression models, for which a normalized index is introduced and a Monte Carlo significance test of statistical hypothesis is provided.

The selection of a subset of variables from a pool of candidates is an important problem in regression, classification, clustering and other areas of multivariate statistics (Woodruff and Reiners, 2004; Brusco *et al.*, 2009; Bartkutė-Norkūnienė, 2009; etc.). Support vector machines (Vapnik, 1995) is a statistical technique that involves the adoption of various heuristic procedures for implementing this technique to classification or regression. Multiclass SVM were introduced by Weston and Watkins (1998) and over more than a decade, many multiclass SVMs have been developed (see, Guermeur, 2010; Balys and Rudzkis, 2010; etc.). Using the SVM one requires to set the values of two types of hyperparameters: the soft margin parameter and the parameters of the kernel. The leave-one-out-one-appears is attractive in a cross-validation procedure applied to selection of the SVM model, because it is known how to produce an estimator of the generalization error which is almost unbiased. Since it is highly time-consuming, a new multiclass SVM with a quadratic loss function has been developed (Guermeur and Monfrini, 2011), in which a generalized radius-margin bound on the leave-one-out error of the hard margin version is established and assessed. That provides us with a differentiable objective

function to perform model selection for the multiclass SVM. A comparative study including all the four multiclass SVMs illustrates the generalization performance of the new machine.

Clustering of objects into groups is usually performed using a statistical heuristic or an optimisation (Bagirov, 2008; Jessop, 2010; Xavier, 2010). The solving method depends on the size of the problem and its purpose. The computational problems arising here belong to the group of NP-hard problems of combinatorial optimization, where the objective function is to find the subset of variables that optimize the value of some goodness of the fit function. The clustering is a relevant approach in detecting communities in networks that is a problem of an increasing importance. The solution of clustering problems in community detection has been studied in Liang and Szeto (2011), where a recursive algorithm is designed to detect a community by clustering nodes intelligently. Thus, the problem of community detection becomes one of the finding groups of connected nodes so that the density of links within a group is the highest one, resulting in an image analysis problem for an adjacency matrix with the highest density contrast. Since the relabelling of network nodes does not alter the topology of the network, the problem of community detection corresponds to the finding of a good labelling of nodes so that the adjacency matrix forms blocks. By putting a fictitious interaction between nodes, the relabelling problem becomes that of energy minimization, where the total energy of the network is defined by putting interaction between the labels of nodes so that clustering nodes that are in the same community will decrease the total energy. A greedy method is used for the computation of minimum energy. The method shows efficient detection of communities in artificial as well as real world networks.

Data simplification and presentation depend to a framework in which the methods of classical statistical analysis are combined with the heuristics ones. Vector quantization is often applied to reduce the amount of data usually forming a quantized approximation to the distribution of the input data vectors, using a finite number of codebook vectors. Once the codebook is chosen, the approximation of data means finding the codebook vectors closest to the initial data vectors (Kohonen, 2001). Thus, the purpose of vector quantization for a given data set is to discover the optimal codebook, containing a predetermined number of codebook vectors, which guarantee the minimization of the chosen distortion metric (usually Euclidean distance) for all the vectors from the data set. Each codebook vector has an associated integer index used for referencing. Vector quantization is used for data compression, missing data correction, classification, etc. It can be used for data clustering, too. In the latter case, the codebook vectors are representatives of clusters. The quality of quantization and visualization of vectors, obtained by two vector quantization methods, namely, self-organizing map and neural gas, have been investigated by Kurasova and Molytė (2011). The number of neuron-winners, quantization and visualization qualities, and preservation of the data structure in the mapping image are explored. The quality of quantization has been measured by a quantization error using numerical measures for proximity preservation (Konig's topology preservation measure and Spearman's correlation coefficient) and a multidimensional scaling for visualization of multidimensional vectors.

When analyzing the measurements coming from real problems, some events or anomalies are often manifested by abrupt changes which may not be quite apparent in their early stage. The representation of vectors in data by collections of parameterized waveforms instead of just representing signals as superpositions of sinusoids ensure the satisfying rate of convergence of representation in the presence of jumps (see, Chen *et al.*, 1998). The Heaviside dictionary merged with the Fourier or wavelet dictionary can solve the problem quite satisfactorily and denoise signal exhibiting discontinuities in the mean (Neubauer and Vesely, 2011). The latter contribution is focused on the change point detection in a one-dimensional stochastic process by sparse parameter estimation from an overparameterized model. The basis pursuit algorithm is used to get the sparse parameter estimates and can be proposed as an alternative to conventional statistical techniques of the change point detection. Unlike them, rather than on stochastic reasoning the basis pursuit relies mainly on geometric ideas where small linear correlations (normalized scalar products) of the noise vector with dictionary atoms are a dominant condition. The change point detection techniques may be useful, for instance, in modeling economic or environmental time series where jumps can occur.

Sometimes large amounts of data are necessary in cognition of stochastic processes depending on unknown parameters. Here we need to find statistics that gives the best estimate of the considered parameter. The paper of Kubilius and Melichov (2011) focuses on finding the optimal estimator for the Hurst index of the fractional Brownian motion and processes based on it.

3. Heuristics in Applications

Markov models are wideused in computer science, statistics, biology, engineering, communication network modelling, etc. The main drawback of Markov chains modelling is a rapid growth of system states. The real engineering and industrial problems can be very complex which causes large Markov chains (having thousands or even millions of states). Thus, the use of heuristic model specification techniques and fast numerical solution algorithms can be succesfuly applied to solve real-life problems. Several nonhomogeneous semi-Markov models have been proposed (D'Amico *et al.*, 2011) as a useful tool for predicting that are implemented in quantification of the effects in HIV evolution due to the age and medical progress. The nonlinear semi-Markov models with respect to standard epidemiologic data analysis have many advantages, because they include the randomness both in the different states of infection and in the time spent in each state, and also take into into account the different ages of patients in the evolution of infection as well as improvements of disease through interrelated states (Corradi *et al.*, 2004; Mathieu *et al.*, 2007). Besides, these models are grounded by few commonly applied and weak hypotheses, and allow us to draw conclusions based on the list of all the computed probabilities descending directly from the observed data through nonparametric models. Indeed, by means of the nonhomogeneous model, it is possible to study a dynamic evolution of the infection differentiated according to the patient's age and scientific medical progresses

due to the elapse of time. In this way, the model could be useful in order to manage the flows of patients towards a hospital.

Service providers and managers are interested in developing criteria for clustering customers into meaningful groups according to their expected length of service. Problems of this kind arising in the health care sphere are analyzed in Garg *et al.* (2011) and two new techniques are presented for clustering the patients' hospital length of stay using phase-type survival trees and mixed distribution survival trees. Key advantage of these new types of survival trees is their ability to accurately model customer pathways, followed by different groups of patients, as a finite state continuous time Markov chain that facilites an easy analysis of models and better explainability to professionals than other survival tree models.

Tools for exploring and modelling the relationships to be predicted between several datasets are often applied in chemometrics, sensometrics and process monitoring. An extension of redundancy analysis in order to improve the fitting ability of multiblock partial least square regression, widely used for exploring and modelling the relationships between several datasets, is given in Stefanie *et al.* (2011). The multiblock approach developed makes it possible not only to combine several sources of information, but also highlights the importance of each data block in the prediction of the response variables. The interests of the multiblock methods are illustrated on the basis of a simulation study and on a real dataset in the field of veterinary epidemiology.

A wide field of application of heuristics is the decision-making related to technology and economics. The problems arising in this field pose the need to deal with substantial amounts of information and developmet of heuristic methods for discrete and continuous optimisation. One of the areas of a wide application of heuristic procedures to decision-making is the Multicriteria Decision Making – MCDM) (Zavadskas *et al.*, 2009; Kaklauskas *et al.*, 2010; Kanapeckienė *et al.*, 2010). Heuristic procedures based on an application of Monte Carlo simulation combined with MCDM allow us making decisions in the presence of uncertain information (Zavadskas and Vaidogas, 2009; Vaidogas and Šakėnaitė, 2010).

4. Large Scale Multidimensional Data Visualization

At present, computer systems store large amounts of data. Data from the real world are often described by an array of features, i.e., we deal with multidimensional data. It is much easier for a human to observe, detect, or extract some information from the graphical representation of these data (to detect the presence of clusters, outliers or various regularities in the analysed data) than from the raw numbers.

A large number of approaches for graphical representation (visualization) of multidimensional data are available. A well-known method is the Principal Component Analysis (Jolliffe, 2002) that provides the mean square optimized linear projection of data. Another classical method is Multidimensional Scaling – MDS (Cox and Cox, 2001; Borg and Groenen, 2005) that works with inter-point distances and gives a low-dimensional

configuration, which represents the given distances best. In all the cases, the visualization problems are formulated as the optimization ones and solved via optimization of heuristically justified objective functions. Neural networks find an application here, too: SAMANN neural network (Mao and Jain, 1995; Medvedev and Dzemyda, 2006), the Self Organizing Map – SOM (Kohonen, 2001), and approaches integrating SOM and MDS (Dzemyda, 2001; Dzemyda and Kurasova, 2006, Kurasova and Molytė, 2011).

There are other visualization methods that take into account the specific character of the analyzed multidimensional data. In most cases the points that contain the important knowledge extracted from real-life data, lie on a low-dimensional manifold embedded in a high-dimensional space. A large number of nonlinear manifold learning methods have been proposed and investigated over the last decade: ISOMAP (Tenenbaum *et al.*, 2000; Han *et al.*, 2009), Locally Linear Embedding (Saul and Roweis, 2003; Karbauskaitė and Dzemyda, 2009a; Karbauskaitė *et al.*, 2010), Laplacian Eigenmaps (Belkin and Niyogi, 2003; Karbauskaitė and Dzemyda, 2009b), etc. Nonlinear manifold learning methods automatically discover the low-dimensional nonlinear manifold in a high-dimensional data space and then embed the data points into a low-dimensional embedding space, preserving the underlying structure in the data.

The MDS method is unsuitable for large data sets: it takes much computing time or there is not enough computing memory. Similar problems arise in neural networks-based visualization and manifold learning. Some ideas for solving the problems arising in the large scale data visualization are reviewed below. Various heuristical modifications of MDS have been proposed to visualize the large data sets: Steerable Multidimensional Scaling – MDSteer (Williams and Munzner, 2004), Incremental MDS, Relative MDS (Naud, 2004; Bernatavičienė, *et al.*, 2007a), Landmark MDS (de Silva and Tenenbaum, 2003), Diagonal Majorization Algorithm – DMA (Trosset and Groenen, 2005), etc. In the paper of Bernatavičienė *et al.* (2007b), the diagonal majorization algorithm has been investigated. The research focuses on the possibilities to increase the efficiency of the algorithm by disclosing its properties. The experiments have proved that, when visualizing a large data set with DMA, it is possible to save the computing time taking into account several factors: the strategy of numbering the multidimensional vectors in the analysed data set and the neighbourhood order parameter. The paper Ivanikovas *et al.* (2008) suggests a way for large datasets visualization with a neural network using clustered training data. It has been noticed that it is possible to speed-up the SAMANN network training process, using a part of the analyzed dataset. The results of the experiments have proved that it is possible to find such a subset of the analyzed dataset that, while training the SAMANN network by it, lower projection errors are obtained faster (more than 10 times) than by training with all the vectors of the set: the usage of the reduced dataset for SAMANN training enabled us to process large datasets and to get good enough results within a reasonable time.

Modern software realizations for large-scale multidimensional data visualization are under the development. The pervasive concept of cloud computing suggests that visualization, which is both data and computing intensive, is a perfect cloud computing application. In the paper of Dzemyda *et al.* (2011), an approach of the web application in

data mining, oriented to the multidimensional data visualization is proposed. This paper focuses on visualization methods as a tool for the visual pattern recognition in large-scale multidimensional data sets. The proposed web service (web application) simplifies the usage of visualization methods that are often very sophisticated. The realization of such a web application receives a multidimensional data set and as a result produces a visualization of the data set. It also supports different configuration parameters of the used data mining methods. These parameters allow the user to control the visualization process and to extract knowledge from their data set much more comprehensively. The paper of Tanahashi *et al.* (2010) presents another sketch of an interface design for the online visualization service.

5. Conclusions

The paper presents a survey of articles, both theoretical and practical, providing new results and new methods potential for solving real-life problems by analyzing the relevant data. The use of recent advances in different fields of large-scale data analysis is promoted in the heuristics framework, focusing on applications in medicine, biology and technology.

References

Bagirov, A. (2008). Modified global k-means algorithm for minimum sum-of-squares clustering problems. *Pattern Recognition*, 41(10), 3192–3199.

Balys, V., Rudzkis, R. (2010). Statistical classification of scientific publications. *Informatica*, 21(4), 471–486.

Bartkutė-Norkūnienė, V. (2009). Stochastic optimization algorithms for support vector machines classification. *Informatica*, 20(2), 173–186.

Belkin, M., Niyogi, P. (2003). Laplacian eigenmaps for dimensionality. *Speech Communication*, 1(2–3), 349–367.

Bernatavičienė, J., Dzemyda, G., Marcinkevičius, V. (2007a). Conditions for optimal efficiency of relative MDS. *Informatica*, 18(2), 187–202.

Bernatavičienė, J., Dzemyda, G., Marcinkevičius, V. (2007b). Diagonal majorization algorithm: properties and efficiency. *Information Technology and Control*, 36(4), 353–358.

Benati, S. (2011). Heuristic methods for the optimal statistic median problem. *Computers and Operations Research*, 38(1), 379–386.

Borg, I., Groenen, P.J.F. (2005). *Modern Multidimensional Scaling*, 2nd edn. Springer Series in Statistics, Springer, Berlin.

Box, G.E.P. (1979). Robustness in the strategy of scientific model building. In: Launer, R., Wilkinson, G. (Eds.), *Robustness in Statistics*. Academic Press, San Diego. pp. 201–235.

Brusco, M.J., Singh, R., Steinley, D. (2009). Variable neighborhood search heuristics for selecting a subset of variables in principal component analysis. *Psychometrika*, 74(4), 705–726.

Chen, S.S., Donoho, D.L., Saunders, M.A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1), 33–61.

Chiyoshi, F., Roberto, D,. Galva, R.D. (2000). A statistical analysis of simulated annealing applied to the p-median problem. *Annals of Operations Research*, 96, 61–74.

Corradi, G., Janssen, J., Manca, R. (2004). Numerical treatment of homogeneous semi-Markov processes in transient case – a straightforward approach. *Methodology and Computing in Applied Probability*, 6, 233–246.

Cox, T.F., Cox, M.A.A. (2001). *Multidimensional Scaling*, 2nd edn., Chapman Hall/CRC, London/Boca Raton.

D'Amico, G., Di Biase, G., Janssen, J, Manca, R. (2011). HIV evolution: a quantification of the effects due to age and to medical progress. *Informatica*, 22(1), 27–42.

De Silva, V., Tenenbaum, J.B. (2003). Global versus local methods for nonlinear dimensionality reduction. In: Becker, S., Thrun, S., Obermayer, K. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 15. MIT Press, Cambridge. pp. 721–728.

Durio, A., Isaia, E.D. (2011). The minimum density power divergence approach in building robust regression models. *Informatica*, 22(1), 43–56.

Dzemyda, G. (2001). Visualization of a set of parameters characterized by their correlation matrix. *Computational Statistics and Data Analysis*, 36(10), 15–30.

Dzemyda, G., Kurasova, O. (2006). Heuristic approach for minimizing the projection error in the integrated mapping. *European Journal of Operational Research*, 171(3), 859–878.

Dzemyda, G., Marcinkevičius, V., Medvedev, V. (2011). Web application for large-scale multidimensional data visualization. *Mathematical Modelling and Analysis* (accepted).

Han, D., Choi, H., Park, C., Choe, Y. (2009). Fast and accurate retinal vasculature tracing and kernel-Isomap-based feature selection. In: *Proceedings of the 2009 International Joint Conference on Neural Networks*, IJCNN'09. IEEE, New York. pp. 1075–1082.

Garg, L., McClean, S., Meenan, B., Millard, P. (2011). Phase-type survival trees and mixed distribution survival trees for clustering patients' hospital length of stay. *Informatica*, 22(1), 57–72.

Guermeur, Y. (2010). Sample complexity of classifiers taking values in RQ, application to multi-class SVMs. *Communications in Statistics – Theory and Methods*, 39(3), 543–557.

Guermeur, Y., Monfrini, E. (2011). A quadratic loss multi-class SVM for which a radius-margin bound applies. *Informatica*, 22(1), 73–96.

Ivanikovas, S., Dzemyda, G., Medvedev, V. (2008). Large datasets visualization with neural network using clustered training data. In: *Advances in Databases and Information Systems. Lecture Notes in Computer Science*, Vol. 5207. Springer, Berlin. pp. 143–152.

Jessop, A (2010). An optimising approach to alternative clustering schemes. *Central European Journal of Operations Research*, 18(3), 293–309.

Jolliffe, I.T. (2002). *Principal Component Analysis*. Springer, Berlin.

Kaklauskas, A., Zavadskas, E.K., Naimavičienė, J., Krutinis, M., Plakys, V., Venskus, D. (2010). Model for a complex analysis of intelligent built environment. *Automation in Construction*, 19(3), 326–340.

Kanapeckienė, L., Kaklauskas, A., Zavadskas, E.K., Seniut, M. (2010). Integrated knowledge management model and system for construction projects. *Engineering Applications of Artificial Intelligence*, 23(7), 1200–1215.

Karbauskaitė, R., Dzemyda, G. (2009a). Topology preservation measures in the visualization of manifoldtype multidimensional data. *Informatica*, 20(2), 235–254.

Karbauskaitė, R., Dzemyda, G. (2009b). Dependence of the Laplacian eigenmaps method and its modification on the parameters. In: *Proc. of the XIIIth International Conference "Applied Stochastic Models and Data Analysis"*, ASMDA-2009. Technika, Vilnius. pp. 263–268.

Karbauskaitė, R., Dzemyda, G., Marcinkevičius, V. (2010). Dependence of locally linear embedding on the regularization parameter. *TOP, An Official Journal of the Spanish Society of Statistics and Operations Research*, 18(2), 354–376.

Kleijnen, J.P.C., den Hertog, D., Angun, E. (2004). Response surface methodology's steepest ascent and step size revisited. *European Journal of Operational Research*, 159(1), 121–131.

Kohonen, T. (2001). *Self-Organizing Maps*, 3rd edn. *Springer Series in Information Sciences*, Vol. 30. Springer, Berlin.

Kubilius, K., Melichov, D. (2011). On comparison of the estimators of the Hurst index of the solutions of stochastic differential equations driven by the fractional Brownian motion, *Informatica*, 22(1), 97–114.

Kurasova, O., Molytė, A. (2011). Quality of quantization and visualization of vectors obtained by neural gas and self-organizing map. *Informatica*, 22(1), 115–134.

Liang, T., Szeto, K.Y. (2011). Community detection through optimal density contrast of adjacency matrix. *Informatica*, 22(1), 135–148.

Mao, J., Jain, A.K. (1995). Artificial neural networks for feature extraction and multivariate data projection. *IEEE Trans. Neural Networks*, 6, 296–317.

Mathieu, E., Foucher, Y., Dellamonica, P., Daures, J.P. (2007). Parametric and nonhomogeneous semi-Markov process for HIV control. *Methodology and Computing Applied Probability*, 9(3), 389–397.

Medvedev, V., Dzemyda, G. (2006). Optimization of the local search in the training for SAMANN neural network. *Journal of Global Optimization*, 35, 607–623.

Meiri, R., Zahavi, J. (2006). Using simulated annealing to optimize the feature selection problem in marketing applications. *European Journal of Operational Research*, 171(3), 842–858.

Michalewicz, Z., Fogel, D.B. (2004). *How to Solve It: Modern Heuristics*. Springer, Berlin.

Naud, A. (2004). Visualization of high-dimensional data using a association of multidimensional scaling to clustering. In: *Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems*, Vol. 1. pp. 252–255.

Neubauer, J., Vesely, V. (2011). Change point detection by sparse parameter estimation. *Informatica*, 22(1), 149–164.

Nunkesser, R, Morell, O. (2010). An evolutionary algorithms for robust regression. *Computational Statistics and Data Analysis*, 54(12), 3242–3248.

Pearl, J. (1983). *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley, Reading.

Sakalauskas, L., Krarup, J. (2006). Heuristic and stochastic methods in optimization. *European Journal of Operational Research*, 171(3), 723–890.

Sakalauskas, L., Zavadskas, E. (2009). Optimization and intelligent decisions. *Technological and Economic Development of Economy*, 15(2), 189–196.

Saul, L.K, Roweis, S.T. (2003). Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4, 119–155.

Stephanie, B., El Mostafa, Q., Lupo Coralie, L., Mohamed, H. (2011). From multiblock partial least squares to multiblock redundancy analysis: a continuum approach. *Informatica*, 22(2) (accepted).

Tanahashi, Y., Chen, C.-K., Marchesin, S., Ma, K.-L. (2010). An interface design for future cloud-based visualization services. In: *Proceedings of 2010 IEEE Second International Conference on Cloud Computing Technology and Science*. IEEE, New York. pp. 609–613.

Tenenbaum, J.B., de Silva, V., Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2319–2323.

Trosset, M.W., Groenen, P.J.F. (2005). Multidimensional scaling algorithms for large data sets. *Computing Science and Statistics*, 37.

Tversky, A.; Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, New Series, 185(4157), 1124–1131.

Vaidogas, E.R., Šakėnaitė, J. (2010). Protecting built property against fire disasters: multi-attribute decision making with respect to fire risk. *International Journal of Strategic Property Management*, 14(4), 391–407.

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer, Berlin.

Williams, M., Munzner, T. (2004). Steerable, progressive multidimensional scaling. In: *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS'04. IEEE, New York. pp. 57–64.

Winker, P., Gilli, M. (2004). Application of optimization heuristics to estimation and modeling problems. *Computational Statistics and Data Analysis*, 47(2), 211–223.

Xavier, A.E. (2010). The hyperbolic smoothing clustering method. *Pattern Recognition*, 43(3), 731–737.

Zavadskas, E.K., Kaklauskas, A., Vilutienė, T. (2009). Multicriteria evaluation of apartment blocks maintenance contractors: Lithuanian case study. *International Journal of Strategic Property Management*, 13(4), 319–338.

Zavadskas, E.K., Vaidogas, E.R. (2009). Multiattribute selection from alternative designs of infrastructure components for accidental situations. *Computer-Aided Civil and Infrastructure Engineering*, 24(5), 346–358.

G. Dzemyda graduated from Kaunas University of Technology, Lithuania, in 1980, and in 1984 received there the doctoral degree in technical sciences (PhD) after postgraduate studies at the Institute of Mathematics and Informatics, Vilnius, Lithuania. In 1997 he received the degree of doctor habilius from Kaunas University of Technology. He was conferred the title of professor (1998) at Kaunas University of Technology. He is a director of the Vilnius University Institute of Mathematics and Informatics and heads the System Analysis Department of the institute. The areas of research are the theory, development and application of optimization, and the interaction of optimization and data analysis. The interests include visualization of multidimensional data, optimization theory and applications, data mining in databases, multiple criteria decision support, neural networks, parallel optimization.

L. Sakalauskas has graduated from the Kaunas Polytechnical Institute (1970), received the PhD degree from this Institute (1974) and the degree of doctor habilius from the Institute of Mathematics and Informatics (2000). President of the Lithuanian Operational Research Society (2010), elected member of the International Statistical Institute (2002), presently is a head of the Operational Research Sector of the Vilnius University Institute of Mathematics and Informatics and professor of the Department of Information Technologies of the Vilnius Gediminas Technical University. His research interests include stochastic modeling and optimization with applications.

Didelės apimties duomenų analizė euristiniais metodais

Gintautas DZEMYDA, Leonidas SAKALAUSKAS

Duomenų analizėje dažnai kyla uždaviniai, kuriuos sunku ar kartais net neįmanoma išspręsti standartiniais skaičiavimo metodais. Straipsnyje apžvelgiamos dažniausiai naudojamos didelės apimties duomenų analizės strategijos, grindžiamos euristiniais sprendimais. Atkreiptas dėmesys į taikymus įvairiose srityse, kur atsiranda uždavinių priimti sprendimus analizuojant didelius kiekius duomenų tame tarpe medicinoje, biologijoje ir technikoje.