INFORMATICA, 2009, Vol. 20, No. 2, 165–172

Optimization and Knowledge-Based Technologies

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

Received: November 2008; accepted: January 2009

Abstract. Recent changes in the intersection of the fields of intelligent systems optimization and statistical learning are surveyed. These changes bring new theoretical and computational challenges to the existing research areas racing from web page mining to computer vision, pattern recognition, financial mathematics, bioinformatics and many other ones.

Keywords: optimization, knowledge-based technologies, statistical learning, data mining, visualization.

1. Introduction

The rapid development of IT has paved the way to a search for techniques capable of dealing with a huge amount of data information around us in all the branches of sciences. Recent few years have seen several changes in the intersection of the fields of intelligent systems optimization and statistical learning:

- the exponential increase in computational power has enabled intelligent systems to process and explore massive data sets;
- the amount of data collected has exploded (Goldman Sach), for example, keeps track of 1.1 million financial time series; (Vault, 2006), similarly as in gene research (Masseroli, 2003);
- new computationally intensive methods for finding information in such data, sometimes called "data mining" techniques, have emerged in a number of fields including neural networks, genetic algorithms, artificial intelligence, machine learning, and data engineering;
- visualization becomes more widely integrated into the knowledge discovery process in search for intelligent decisions – visual data exploration strives for integrating humans into the data exploration process, by applying their perceptual abilities of large data sets, that are available at present.

These changes bring new theoretical and computational challenges to exciting research areas ranging from web page mining to computer vision, pattern recognition, financial mathematics and bioinformatics, etc. The support vector machines, margin classifiers, k-means clustering, perceptron algorithm, are just a few examples of the terminologies we often encounter in classification, clustering and machine learning

problems. Optimization techniques pierced through many such methodologies, and a variety of other properties, leading to interpretation and computation of solutions, relies on optimization techniques as well. The scientific aim of this issue consists in seeking solutions to challenging intelligent decision problems within the use and development of advanced/modern mathematical optimization and statistics tools. The special structures often encountered in data mining systems also provide motivation for further investigations on new optimization methods and algorithms capable of handling very large scale data, and hence this issue will advance both areas.

2. Optimization and Data Mining

Clustering, considered as the most important unsupervised learning problem, deals with the partitioning of a data set to groups of similar item (Hand *et al.*, 2001). The cluster validation problem and determination of the true number of clusters are very important in clustering (Barzily *et al.*, 2009). Stability-based approaches for the cluster validation problem, evaluate the variability of partitions under repeated applications of a clustering algorithm. The low variability is understood as the high consistency in the results obtained (see, for example, Cheng *et al.*, 1996). Therefore, the number of clusters that maximizes the cluster stability is accepted as an estimate for the "true" number of clusters in a data set. The approach that the clusters correspond to modes of the probability density function of the data elements is adopted in the following way. Pairs of samples are drawn from the data, according to two sampling distributions. The first distribution corresponds to the high density zones of data-elements distribution. Thus, it is associated with the cores of clusters. The second one, associated with the cluster margins, is related to the low density zones. The samples are clustered and the two partitions obtained are compared. The partitions are considered to be consistent, if the obtained clusters are similar. The resemblance is measured by the total number of edges, in the minimal spanning trees of clusters, connecting the points from different samples. Following the conclusion that the true number of clusters corresponds to the statistic empirical distribution, which is closest to normal, the Friedman and Rafsky two sample test statistic is implemented for the analysis of clustering stability (Fridman *et al.*, 1979). Numerical experiments corroborated the ability of the approach to detect the true number of clusters.

The common data mining technique for finding hidden patterns in data is the classification analysis. Support Vector Machines (SVMs) are an efficient approach for classifying as well as for clustering, time series, and regression analysis (Cortes and Vapnik, 1995). Very often a part of the available data is unlabelled, which implies the semi-supervised classification problem, that solved by Support Vector Machines (SVM), and explored as an unconstrained and non-smooth optimization task. The objective function of such problems is a mathematical expectation functional depending on decision rules (or strategies), i.e., on functions of the observed random parameters. This kind of problems arises in on-line decision making and in the statistical learning theory. The problem is solved by sample average approximation combined with Tikhonov's

regularization technique (Norkin and Keyzer, 2009). In view of presented theorem the solution of the approximate regularized problem is searched in a kernel form, i.e., as a finite linear combination of kernel functions, and thus, it is found through a finite-dimensional optimization. Non-asymptotic bounds on the deviation of the expected risk of the approximate solution are derived from its least possible value and sufficient conditions for an uniform convergence of kernel solutions are established, jointly with the rule for downward adjustment of the regularization factor with an increasing sample size. As special cases, the asymptotic convergence of kernel regression estimators and binary classifiers, as a sample size tends to infinity, has been studied, too.

The main goal of classification problems is to formulate a criterion (objective function) in an nalytical form, which allows us to visualize data in a two-dimensional visualization space such as the classification rule of multidimensional data established by SVM to satisfy the classification rule of the two-dimensional data visualized. If the classification results could be directly seen in a two-dimensional space, it will be a great help for classification users. High dimensional data and binary classification results can be visualized in a low dimensional space. A simple and direct idea for visualizing high-dimensional data is to reduce its dimensionality to two or three dimensions by using some dimension reduction algorithms. Feature extraction is the process of mapping the original features (measurements) into fewer features, which preserve the main information of the data structure. Feature extraction for exploratory data projection enables us to visualize high-dimensional data for a better data structure understanding and for cluster analysis. A large number of approaches to data projection (mapping) 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 that represents the given distances best. In all the cases, the mapping problems are formulated as the optimization ones. Neural networks find an applicaion here, too: SAMANN neural network (Mao and Jain, 1995; Medvedev and Dzemyda, 2006), the Self Organizing Map (SOM; Kohonen, 2001), approaches integrating SOM and MDS (Dzemyda, 2001; Dzemyda and Kurasova, 2006).

There are other visualization methods that take into account the specific character of the analyzed multidimensional data. A large number of nonlinear manifold learning methods have been proposed over the last decade: Locally Linear Embedding (Saul and Roweis, 2003), Laplacian Eigenmaps (Belkin and Niyogi, 2003), etc. At most cases the points, containing the important knowledge extracted from real-life data, lie on a low-dimensional manifold embedded in a high-dimensional space. 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 paper of Karbauskaitė and Dzemyda (2009) deals with the optimization of quality of the locally linear embedding. In order to quantitatively estimate the topology preservation of a manifold after unfolding it in a low-dimensional space, some quantitative numerical measures are introduced and investigated. Three measures have been investigated:

Spearman's rho, Konig's measure (KM), and mean relative rank errors (MRRE). After investigating different manifolds, it turned out that only KM and MRRE gave better results of manifold topology preservation in all the cases.

Stochastic optimization is a way of avoiding difficulties related with convexity and non-differentiability of the objective function in the semi-supervised classification, solved by means of SVM and mathematical programming. Applicability of simulated annealing (Yang, 2000) and stochastic approximation (Ermoliev, 1976) methods to the classification by SVM has been explored by computer simulation (Grigaitis *et al.*, 2007; Bartkutė-Norkūnienė, 2009). The numerical simulation studies with several synthetic and real-life datasets show that these algorithms can be successfully applied to optimizing non-differentiable loss functions in the classification problems. On the base of the simulation results a conclusion can be drawn that simulated annealing has higher sensitivity, specificity, and overall accuracy rates as compared with the stochastic approximation algorithms. The main advantage of the methods proposed is the possibility to train the classifier on the basis of a large number of labelled and unlabelled points. This could be useful especially in many real-life classification problems.

3. Optimization and Complex Data Analysis

A lot of optimization problem are closely connected with the problem of optimal learning in the class of neural perceptrons. One such problem is the minimum affine separating committee (MASC) conbinatorial optimization (Khachay and Pobery, 2009). A special interest is caused with the results regarding estimations of computational complexity of the learning problem by the elementary networks, namely, classical perceptrons. Computational complexity of the problem of training a layered neural network, being optimal according to one or another criteria, interests researchers since the 1980's and it is known that this problem is NP-hard (Judd, 1987; Lin and Vitter, 1991). But the question whether this problem, being formulated in a feature space of the fixed dimensionality is NP-hard as well, or it has a polynomial-time algorithms in this case, is still open, while it is known, that the MASC problem in an one-dimensional space belongs to P class of complexity. Thus, the intractability of the MASC, problem in an n-dimensional space with fixed $n > 1$ has been proven. Actually, it is shown that the MASC, being formulated in a fixed-dimensional feature space, is intractable even if the sets that define the instance were in the general position. Finally, a new approximation algorithm for the MASC problem in the fixed dimension feature space has been presented.

Intersection of different scientific fields provides efficient solutions for intelligent decision making very often (Sakalauskas, 2006). An example of this is the scheduling problem solved by combining discrete-time nonlinear optimal control of dynamic system models with state and control continuous flow variables subjected to upper and lower bounds (Fagundez *et al.*, 2009). Complementarity constraints are used to represent scheduling decisions in this approach. All constraints are modeled without discrete variables, achieving a continuous model. These models are smaller than their

mixed integer linear programming counterparts, known from the literature. Although being capable of generating good solutions, the nonlinear programming formulation is nonconvex, however, it is employed as an auxiliary problem in the traditional mixed integer linear programming formulation. In fact, preliminary numerical results showed a significant reduction of iterations when initialized by the solution of such an auxiliary problem.

The knapsack problem occurs in many applicable fields, besides being a counterpart in many combinatorial and discrete optimization problems. The performance of the modified subgradient algorithm (MSG) to solve the 0–1 quadratic knapsack problem has been examined (Sipahioglu and Sarac, 2009). The MSG has some important proven properties. For example, it is convergent, and it guarantees a zero duality gap for the problems such that their objective and constraint functions are all Lipschitz (Gasimov, 2002). Besides, the MSG has been successfully used for solving nonconvex continuous and some combinatorial problems with equality constraints since its suggestion. Using the MSG to solve the qudratic knapsack problem, a zero-one nonlinear problem is converted into a continuous nonlinear problem by adding only one constraint and not adding new variables. Then, to solve the continuous QKP, a dual problem with a "zero duality gap" is constructed by using the sharp Augmented Lagrangian function. Finally, the MSG is used to solve the dual problem, by considering the equality constraint in the computation of the norm. To compare the performance of the MSG, some literature test instances were solved both by using the MSG as well as some GAMS' and MINLP solvers.

Analysis of complex data is often reduced to computational problems, solved by combining user-given black-box procedures – usually heuristics themselves – in the hope of obtaining a more efficient or more robust procedure. Heuristics are generally applied to problems for which there is no satisfactory problem-specific algorithm. Genetic algorithms, modeling evolutionary processes are commonly used heuristics (Goldberg, 1989). In the paper of Misevičius and Rubliauskas (2009), several Hybrid genetic algorithms (HGAs) for solving structured quadratic assignment problems (QAP) have been investigated, in partucular the following ones: compounded approach, incorporating the expensive local search, reinforced improvement, fitness-based selection, gender modification, cohesive crossover, maintaining many offspring, elitism strategy, and using restarts. The results of the computational experiments with the structured QAP instances taken from the electronic library QAPLIB demonstrate the significant benefit of the proposed modifications. This is especially true for the compounded approach and the entropy-based restart technique. The experiments with these variants indicate that it is of great importance to have high-quality populations and make use of the proper restart mechanism for avoiding the loss of diversity and the premature convergence of the genetic algorithm. The other observation is that very compact populations are enabled which allow saving both the computational time and memory resources. These small-sized populations are fully compensated by including the computationally expensive, but effective local search algorithm and producing a suitable large number of the offspring. The results. It should be noted that the efficiency of algorithms proposed might be improved even more by the careful juxtaposing of the proposed modifications. This

could be one of the future research directions. It may also be worthy to apply these modifications to other combinatorial optimization problems.

The multi-attribute analysis is a useful tool for many intelligent decision-making problems in economy, management, construction, etc. Such decision-making problems are encountered under various situations where a number of alternatives and actions or candidates have to be chosen, based on a set of criteria or attributes (Hwang and Yoon, 1981). The COPRAS (the multiple criteria COmplex PRoportional ASsessment of alternatives) method is widely used for the multi-attribute analysis, while, the accuracy of performance measures in this method is usually assumed to be accurate (Zavadskas *et al.*, 2009). This method assumes a direct and proportional dependence of the weight and utility degree of the investigated versions on a system of attributes adequately describing the alternatives and on values and weights of the attributes. However, usually some uncertainty exists here involved in all multi-attribute model inputs. In real life multi-attribute modelling of attribute values of multi-alternative assessment problems, which deal with the future, can be expressed in intervals. COPRAS-G is a newly developed method for assessment of alternatives by multiple-attribute values determined in intervals. This approach is intended to support the decision making process and increase the efficiency of the resolution process. The COPRAS-G method found many applications in the solution of wide range discrete multi-attribute assessment problems of construction.

4. Discussion and Conclusions

Many algorithms of intelligent data analysis distinguish itself by intractability and complexity because of numerical counterparts related with optimization. In surveyed issue several novel methods are developed and explored being on the intersection of combinatorial and continuous optimization with knowledge-based technologies. Such investigations are especially useful in many real-life problems dealing with large volumes of data.

References

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

Barzily, B., Z. Volkovich, B. Akteke-Öztürk, G.-W. Weber (2009). On a minimal spanning tree approach in the cluster validation problem. *Informatica*, **20**(2), 187–202.

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

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

Cheng, R., *et al.* (1996). Measuring the influence of individual data points in a cluster analysis, *Journal Classification*, **13**, 315–335.

Cortes, C., and V. Vapnik (1995). Support-vector networks. *Machine Learning*, **20**(3), 273–297.

Cox, T.F., M.A.A. Cox (2001). *Multidimensional Scaling*, 2nd ed. Chapman and Hall/CRC.

Deng, J.L. (1988a). Introduction to grey system theory. *The Journal of Grey Theory*, **1**, 1–24.

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

Ermoliev, Yu.M. (1976). *Methods of Stochastic Programming*. Nauka, Moscow (in Russian).

Fagundez F., A. Xavier and J.L.D. Faco (2009). Continuous nonlinear programming techniques to solve scheduling problems. *Informatica*, **20**(2), 203–216.

Friedman, J., *et al.* (1979). Multivariate generalizations of the Wolfowitz and Smirnov two-sample tests, *Annals of Statistics*, **7**, 697–717.

Gasimov R. (2002). Augmented Lagrangean duality and nondifferantiable optimization methods in nonconvex programming. *Journal of Global Optimization*, **24**, 187–203.

Goldberg, D.E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Kluwer Academic Publishers, Boston, MA.

Grigaitis, D., V. Bartkutė, L. Sakalauskas (2007). An optimization of system for automatic recognition of ischemic stroke areas in computed tomography images. *Informatica*, **18**(4), 603–614.

Hand, D., H. Mannila and P. Smyth (2001). *Principles of Data Mining*. The MIT Press.

Hwang, C.L., and K. Yoon (1981). *Multiple Attribute Decision Making – Methods and Application*. Springer-Verlag, Berlin.

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

Judd, J.S. (1987). Complexity of connectionist learning with various node functions. COINS. *Technical Report*, No. 87-60. University of Massachusetts.

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

Khachay, M., and M. Pobery (2009). On the complexity and approximability of committee polyhedral separability. *Informatica*, **20**(2), 217–234.

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

Lin, J.H., J.S. Vitter (1991). Complexity results on learning by neural nets. *Machine Learning*, **6**, 211–230.

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

Masseroli, M., P. Cerveri, P.G. Pelicci, M. Alcalay (2003). GAAS: gene array analyzer software for management, analysis and visualization of gene expression data. *Bioinformatics*, **19**(6), 774–775.

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

Misevičius, A. and D. Rubliauskas (2009). Testing of hybrid genetic algorithms for structured quadratic assignment problems. *Informatica*, **20**(2), 255–272.

Norkin, V., and M. Keyzer (2009). On stochastic optimization and statistical learning in reproducing kernel Hilbert spaces by SVM. *Informatica*, **20**(2), 273–292.

Sakalauskas, L. (2006). Modelling and simulation of business systems. *European Journal of Operational Research*, **175**(3), 1339–14330.

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

Sipahioglu, A., and T. Sarac (2009). The performance of the modified subgradient algorithm on solving the 0–1 quadratic knapsack problem. *Informatica*, **20**(2), 293–304.

Yang (2000). Convergence of the simulated annealing algorithm for continuous global optimization. *Journal of Optimization Theory and Applications*, **104**(3), 691–716.

Vault (2006). *Vault Employer Profile. Goldman Sachs*. New York, Vault, Inc.

Zavadskas, E.K., A. Kaklauskas, Z. Turskis, J. Tamošaitienė (2009). Multi-attribute decision-making model by applying grey numbers. *Informatica*, **20**(2), 305–320.

G. Dzemyda graduated from Kaunas University of Technology, Lithuania, in 1980, and in 1984 received there the doctoral degree in technical sciences (PhD) after post-graduate 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 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 optimization theory and applications, data mining in databases, multiple criteria decision support, neural networks, parallel optimization, internet databases, the models of epidemic spread.

L. Sakalauskas has graduated from the Kaunas Polytechnical Institute (1970), received the PhD degree from this Institute (1974) and the degree of dr. habil. from the Institute of Mathematics and Informatics (2000). Vice-president of Lithuanian Operational Research Society (2001), elected member of the International Statistical Institute (2002), presently is a head of the Operational Research Division of the 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.

Optimizavimas ir žinių technologijos

Gintautas DZEMYDA, Leonidas SAKALAUSKAS

Straipsnyje apžvelgti ir susisteminti pastarųjų metų pokyčiai intelektualiųjų sistemų optimizavimo ir statistinio apsimokymo tyrimų sankirtoje. Šie pokyčiai atveria naujas teorines ir skaičiuojamąsias galimybes žiniatinklio duomenų analizėje, kompiuterinėje vaizdų analizėje, šablonų atpažinime, finansų matematikoje, bioinformatikoje ir daugelyje kitų sričių.