M Ű E G Y E T E M 1 7 8 2

Budapest University of Technology and Economics

Department of Measurement and Information Systems

József Valyon

Extended LS-SVM for System Modeling

PhD thesis

Advisor:
Gábor Horváth PhD
Budapest University of Technology and Economics
Department of Measurement and Information Systems

Budapest, 2007. June

© 2007 József Valyon
Budapest University of Technology and Economics
Department of Measurement and Information Systems
H-1117 Budapest, XI. Magyar Tudósok körútja 2.
Building I. Wing E., Room 329.
Tel: (1) 463 3587, Fax: (1) 463 4112,
Email: valyon@mit.bme.hu

This work was partly sponsored by the Hungarian Fund for Scientific Research (OTKA) under contract T046771

TABLE OF CONTENTS

1. INTRODUCTION

System- identification and modeling is an important way of investigating and understanding the world around. There are several methods that may be used for this task, depending on the amount and type of the information available about the system. If the inside structure of the system is known, a system identification problem can be solved, where the main task is to determine the parameters of known elements. When only input–output observations are obtained, a behavioral or black box model can be constructed. This is an ill-posed problem, where the dataset is usually corrupted by noise and it is not enough to provide a full, detailed description of the system.

This research was conducted at the Budapest University of Technology and Economics, in Department of Measurement and Information Systems, where system identification- and modeling problems have a long history. In conjunction with the Vrije Universiteit Brussel two main research topics are investigated, the frequency domain identification of linear systems [1] and identification of systems containing only small nonlinearities [2]. Black-box modeling problems using intelligent systems, especially neural networks are also subject of extensive research [3]-[5]. A typical application is steelmaking with a Linz-Donawitz (LD) converter [6], which was used to conduct the real-life industrial tests of the proposed methods. This steelmaking problem was solved during a four-year long project ending in 2000. Partly in connection with this project –also with the participation of Vrije Universiteit Brussel - a PhD dissertation was written [7].

The most commonly used kernel-based method is the Support Vector Machine (SVM) [8], propagated by Vladimir Vapnik et al. Recently kernel machines and Support Vector Machines (SVM) have been gaining more and more attention, because they incorporate some useful features that make them favorable in handling the above described situations. As a result of this, kernel methods have become a separate, representative topic of traditional intelligent systems which is currently subject of intensive research.

Kernel- methods, and machines work by applying a nonlinear transformation map the complex, nonlinear primal problem -namely the input samples- to a (finite dimensional) kernel space, where it can be handled linearly. The most commonly used kernel based method is the Support Vector Machine (SVM). The primary advantage of SVM methods is that the constructed model (network) guaranties an upper bound on the generalization error. Another important benefit of applying SVM, is that it aims at constructing a small and compact model, by selecting some vectors -called Support Vectors (SV)- that provide the basis of the model. Since the size of the model corresponds to the number of SVs and usually this number is much less than the number of training samples, the resulting model is sparse. This means that the model complexity or -using the neural interpretation of SVM- the number of neurons is reduced to meet the intrinsic complexity of the problem. This reduction may result in the performance degradation of the model, but the trade-off between performance and sparseness can be adjusted by hyperparameters.

The main drawback of traditional SVM is its high computational complexity, which makes it infeasible for large real-life problems. There are several solutions to overcome this problem. These solutions are mainly iterative "chunking" methods that decompose the large optimization task -involved by SVM construction- to a series of smaller calculations. These methods mainly differ in the way the subtasks, thus the smaller optimization problems are constructed [9].

Another possibility is to change the optimization problem, to achieve a computationally more effective problem. This approach is used in the Smooth Support Vector Machine (SSVM), in its reduced version, the Reduced Support Vector Machine (RSVM) [10] and in the Least Squares SVM (LS–SVM) [11], which provides the basic background of my work. The LS-SVM method was introduced by Johan Suykens. In this case, due to small changes applied the problem (e.g. the use of a Least Squares error criteria), training means solving a set of linear equations instead of the quadratic programming problem involved by the standard SVM. On the other hand the model resulting from the LS-SVM method consists of exactly the same number of kernels as many training samples were used, thus the result is not sparse. By applying a pruning method, the desired sparseness can also be reached by LS-SVM, but in this case the computation time is increased, while performance declines. Another solution is to create a Fixed Size LS-SVM, where the number of support vectors can be defined prior to training.

LS-SVM has also been extended, to reduce the effects of noise, especially of outliers, which is very important concerning real life applications. This method is called Weighted LS-SVM [11].

My work involves system modeling based on real-life, measured –thus usually noisy- datasets, where I focus on classification and function approximation tasks [11].

2. GOALS AND PRACTICES

There are many different modeling approaches that can be applied, which differ in many aspects of the problem, model construction and the result. The most important properties of such solutions can be summarized as follows:

- The complexity of the resulting model.
- The performance (precision) of the model.
- The computational complexity of the modeling method.
- The number free parameters and open questions (e.g. stopping criteria) involved.

The primary goal of the research was to create a kernel-based (SVM) model for the steelmaking problem. Since neither of the available (SVM and LS-SVM) methods fulfills the criteria above, I had to develop a new solution that can combine the desirable features of these methods.

Real-life problem usually comprise huge datasets, which still under represent the modeled system. This means that it is extremely important, to create a method, where the complexity of the resulting model is independent –unlike traditional LS-SVM – of the training set cardinality, but at the same time it should incorporate all available information to provide a good quality model.

In order to provide a precise solution, it is crucial for the method to be able to handle noisy data. The primary goals of this work concerning noisy datasets include the cases of Gaussian noise, noise with known properties (amplitude, distribution) and outliers.

To facilitate the modeling task, one must provide methods, that are computationally effective, in order to use them on complex problems, and large datasets.

A common task of most modeling methods is the selection of the proper (optimal) hyperparameters, which can be a hard and lengthy process. This means that by reducing the number of these parameters or easing the proper selection can simplify the problem.

The goal of this research was to create an SVM based method, that provides a sparse, robust, „easy to build" and computationally effective solution. This is done by extending the LS-SVM formulation with sparseness and robustness.

3. RESEARCH METHODS

The research techniques used in this Thesis come from general backgrounds and theory of system modeling, function approximation, linear algebra, statistics, neural networks and optimization.

This work focuses on eliminating the drawbacks of LS-SVM therefore the thesis are mainly derived from the formulation of this method. In practice, this means that the kernel space representation of the problem, thus the linear equation set involved by LS-SVM construction is used as another, equivalent representation of the primal problem. The reduction and optimization steps are based on this formulation. Similar methods from literature mostly reformulate the primal problem or -the linear, but in practice not calculated- feature representation of the problem. Due to this, the method always orders an obligatory solution method to the kernel space problem, while the proposed approach allows many different solutions originating from this representation.

Another important advantage of using the kernel representation is that it is exact, available (since it is formulated in practice), constructed and linear, so it creates a simple, easy to handle starting point for any extension, modification.

As a result of the above described approach, the propositions work backwards from the kernel space. First they operate on the LS-SVM linear equation set of the kernel space, than the interpretation of these changes are traced back to the feature- and primal space.

To qualify the result one must always consider several properties jointly. The most important aspects of the design:

- The quality of the resulting model.
- The complexity (size) of the model.
- The computational complexity of the method.
- The difficulty level of the implementation.

In the thesis I examine and consider these aspects of the proposed methods.

The mathematical, theoretical backgrounds and the motivations behind a certain method can be provided in most cases, but due to the general complexity („ill-posed") of the problems, there are no exact mathematical tools to evaluate the results. To overcome this problem, benchmark problems and simulations based on cross validation are used.

To evaluate the proposed methods, I have used the most common and widely used benchmark problems, based on the literature of this field.

The real-life industrial modeling task of steelmaking involves regression, therefore I focused on regression methods during my work. Unfortunately, the common regression benchmarks are toy problems, usually used only to illustrate the methods. On the other hand, the classification benchmarks include many complex problems. Since my propositions can be applied to both regression and classification, I have also implemented the algorithms for the classification case, to evaluate the propositions on these benchmarks. To test the regression I have used the simple illustrative samples, defined a more complex time series prediction task and applied it to the LD steel converter modeling.

For the tasks not included in the goals of the research, but involved by the modeling (e.g. hyperparameter selection, kernel selection) I used the most common values and methods.

The testing was done using a my own MATLAB toolbox, which is designed to support all aspects of the extended LS-SVM. Besides incorporating the reduced equation set implementation, it provides a „plug-in" like structure to support many different support vector selection methods (corresponding to thesis 2) and solution methods (summarized by thesis 3).

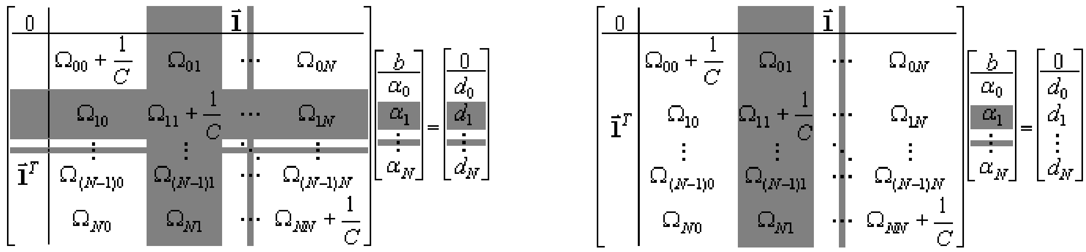
4. NEW SCIENTIFIC RESULTS

Te first Thesis proposes a special "partial reduction" technique, where the LS-SVM training equation set is reformulated to describe a sparse, but precise model and algorithmically more effective problem. In Thesis 2 some method are proposed to support the reduction, thus to select the support vectors. It is also shown, that traditional methods can also be used in conjunction with the partial reduction method. The reduced equation set can be solved using different techniques to achieve more general, e.g. robust estimates. These possible solution methods are summarized by Thesis 3.

1. **Thesis: I have developed a new partially reduced LS-SVM formulation, which provides a sparse solution (reduced model complexity), but still uses all training samples to construct the model; therefore the reduction infers minimal performance degradation, compared to the unreduced (full sized) LS-SVM.**

   In order to achieve a sparse LS-SVM, I have developed a partial reduction method that incorporates all the training samples –all available information- in the learning process, while the model is based on only a subset of all samples, namely the support vectors. The primary aim of LS^2-SVM is to reduce the complexity of the resulting model, thus the number of hidden layer nonlinear neurons. An additional gain of the method is that the model construction becomes algorithmically more effective.

   To achieve a sparse LS-SVM, the traditional pruning method omits some training samples (ones that are considered to be less important) and solves a new smaller quadratic equation set. The partial reduction, however, retains all the training samples, but only a part of them is used as kernel centers. In the main equation set of LS-SVM model building (see Figure 1), the support vectors correspond to columns, while the constraints imposed by the training samples are represented by the rows. It is easy to see, that partial reduction leads to an overdetermined equation set.



a.) b.)

Figure 1. The fully (a.) and the partially (b.) reduced LS-SVM. The gray elements are removed from the equation set.

The omission of columns with keeping the rows means that the network size is reduced; still all the known constraints are taken into consideration. This is the key concept of keeping the quality, while the number of neurons (kernels) is reduced.

The partially reduced equation set (in case of Gaussian noise) is solved in a least-squares sense, thus the method is named Least Squares LS-SVM or shortly LS^2-SVM. This solution can be matched to the traditional LS-SVM, and it is easy to see that in case of no reduction, the two solutions are equivalent.

1.1. I have shown that partial reduction leads to a sparse LS-SVM model, but at the same time preserves all known information (constraints), provided by the training dataset, to construct the solution.

1.2. I have shown that the sparse solution constructed by partial reduction –since all samples are considered- can perform better, than the model constructed by traditional pruning, which entirely omits some samples. On the other hand, the same performance can be achieved by a smaller model.

1.3. Simulations show that in case of partial reduction the effect of regularization is changed, because it is not the only term responsible for avoiding overfitting (thus to achieve good generalization).

1.4. I have shown that the partially reduced LS-SVM, which is derived from the original LS-SVM formulation, corresponds to a model that applies the regularization in the kernel space (instead of the feature space).

2. Thesis: I have proposed new methods that can be used to select the support vectors of the partially reduced model and I have shown that other known methods can also be utilized –with only slight changes– for this task.

In order to create a sparse model based on partial reduction, the support vectors must be determined. There are several methods available for this task, which not only differ in the performance of the resulting model, but in many other aspects, such as computational complexity etc.. The number of support vectors needed to solve a given problem must also be answered. I propose a support vector selection method, which is based on constructing an approximate basis of the kernel space, which vectors are then used as support vectors. The proposed method reduces the differences between LS-SVM and SVM, since the number of support vectors –similarly to the SVM- is controlled by an adjustable tolerance value, determines the trade-off between performance and the degree of sparseness. I have shown that a slight modification of both traditional pruning used to produce a sparse LS-SVM and the Feature Vector Selection (FVS) method proposed for the Reduced Rank Kernel Ridge Regression (RRKRR) [12] can also be used with partial pruning. I analyze and compare the proposed the known methods. Based on traditional pruning, another new selection method is introduced that performs better in certain problems. The statements of thesis 1 and 2 have been published in [s1],[s2],[s3],[s5],[s8], [s10],[s12],[s13],[s17],[s18].

2.1. I have proposed a new effective support vector selection method. Similarly to the traditional SVM, this algorithm incorporates a tolerance parameter which controls the tradeoff between the model size (sparseness) and performance. The proposed method is algorithmically effective compared to the known –mostly iterative– methods. Using simulations, I have shown that this method leads to good results concerning model performance.

2.2. I have shown that previously published selection methods (such as FVS, traditional LS-SVM pruning etc.) can be applied to select the support vectors with only a slight change.

2.3. I have shown that in correspondence with the traditional Fixed Size LS-SVM, partial reduction can also be utilized to create a fixed size LS-SVM model, whilst the good properties of this method are preserved.

2.4. I have shown that a new inverse pruning method can be achieved by using the opposite of the selection criteria used by the traditional pruning method. Using simulations, I have shown that in case of certain hyper parameter settings this method leads to better results than the traditional pruning method.

3. Thesis: I have shown that since the partially reduced LS-SVM leads to an overdetermined system –in the kernel space- the problem can be further analyzed, and optimized solutions (especially ones that reduce the effects of noise) can be found.

After applying the nonlinear transformations in a traditional LS-SVM, only one linear solution can be found in the kernel space, since it's dimensionality equals to the number of mapped data samples. In case of partial reduction, however, the number of training samples exceeds the dimensionality of the problem. This leads to an overdetermined equation set, which has several solutions depending on the optimality criterion used. The most straightforward solution is to minimize the least-squares error (which corresponds to the traditional LS-SVM), but by analyzing the statistical properties of the training data, other optimal solutions may be found. This means that in case of noise, especially outliers, robust models can be calculated. There are solutions described for handling problems with known Gaussian noise and to reduce or even completely remove the effects of outliers.

In handling noisy datasets, different linear solutions are constructed in the kernel space, but a more general approach is also presented by allowing locally linear or even nonlinear solutions, leading to more complex (e.g. multi-layer) models. The robust regression method has been published in refs. [s2],[s3],[s12],[s13].

3.1. I have shown that, if the noise corrupting each training sample is known, than a weighted least-squares solution can provide the optimal solution.

3.2. I have shown that traditional pruning and weighting – although their goals do not rule out each other – cannot be used at the same time, because they work in opposition. The extended LS-SVM is capable of achieving both goals (thus a sparse and robust solution) simultaneously.

3.3. I have shown that applying linear regression methods of robust statistics in the kernel space can reduce, or remove the effects of noise, especially the effects of outliers.

3.4. I have shown that the approximation problem in the kernel space can be further generalized, by allowing locally linear or even nonlinear solutions. The locally linear models permit incremental learning, while the nonlinear solution, in the case when an LS-SVM or SVM is used for the kernel space regression, can lead to a multi layer LS-SVM

5. UTILIZATION OF THE RESULTS

Nowadays there is a large number of industrial modeling problems that need to be solved. Just as described for the LD converter based steelmaking problem, most of these problems require a black-box model to be built based on huge datasets, containing mostly measured, thus noisy data. The propositions described by this Thesis make it possible to apply kernel methods to these problems. The goal is to create a sparse and robust model, with acceptable computational complexity.

The first and second Thesis can be used in conjunction to reduce or overcome the main drawback of traditional LS-SVM, thus to produce a sparse, but precise model of the system. The third Thesis provides solutions to achieve a robust result, in case of noisy datasets especially ones containing outliers.

The usefulness and applicability of the proposed methods are illustrated on widely used benchmark problems, and also on a real-life industrial problem. Steelmaking with an LD converter is a very complex –due to the harsh industrial environment, and the used technology- physical-chemical process, which operates is a dynamic system burdened by large noise (e.g. contaminations etc.). The whole process takes place on very high temperature, in a closed system, therefore there is little, practically no information about the inside process, thus only the input and output measurements can be used.

The extended LS-SVM was successfully applied to this problem:

- Using partial reduction, I have created a small, sparse model for the steelmaking problem, whilst all samples of the large –but from the viewpoint of representing the process still small- dataset has been used. The performance of the model is similar to the results achieved earlier by neural networks.
- I have shown, that by using robust methods, the effects of noise –especially outliers- corrupting the LD converter dataset can be reduced.

Based on the results, a large extent of sparseness can be achieved using the partial reduction method, without significant loss in performance. It can be stated, that by applying this method such small models can be constructed, which if traditional pruning was applied would result in an –intolerably- large error. Experiments show, that the RREF method proposed for selecting support vectors generally provide a good selection, with additional advantages, such as adjustable trade-off between performance and model size etc. The methods used for creating robust models have also been proven very effective, especially in reducing or removing the effects of outliers.

The proposed extensions of the LS-SVM has been published in journals and presented at conferences. It must be mentioned, that the basic idea behind the extended LS-SVM, the LS^2-SVM has been presented in a book titled „Neural networks" which was published in 2006. This book is used as the course material for students majoring in intelligent systems at the Budapest University of Technology and Economics.

Traditional methods usually use the primal, or -the already linear- feature space to search for the desired (sparse, robust etc.) solutions. The propositions of this thesis however use the kernel space representation of the problem, thus the reduction and optimization methods used in the kernel space, offer a wide range of research opportunities.

6. PUBLICATIONS

6.1. Books

[s1] Altrichter Márta, Horváth Gábor, Pataki Béla, Strausz György, Takács Gábor, **Valyon József**, „Neurális hálózatok”, Panem, 2006. (in Hungarian)

6.2. Periodicals

[s2] **J. Valyon** and G. Horváth, „A Robust LS–SVM Regression”, *International Journal of Computational Intelligence, Vol. 3 No. 3.,* pp. 148-153, 2006.

[s3] **J. Valyon** and G. Horváth, „Extended Least Squares LS–SVM”, *International Journal of Computational Intelligence, Vol. 3 No. 3.,* pp. 234-242, 2006.

[s4] **J. Valyon** and G. Horváth, „A Weighted Generalised LS–SVM”, *Periodica Polytechnica Electrical Engineering*, 47/3-4, pp. 229-252., 2003.

[s5] **J. Valyon** and G. Horváth, „A Sparse Least Squares Support Vector Machine Classifier”, *Periodica Polytechnica Electrical Engineering*, pp. 17-23., 2004.

[s6] **Valyon József** és Horváth Gábor, „Least squares szupport vektor gépek adatbányászati alkalmazása”, *Híradástechnika*, pp 33-39, 2005. (in Hungarian)

6.3. International conferences

[s7] P. Berényi, **J. Valyon** and G. Horváth: „Neural Modeling of an Industrial Process with Noisy Data”, *IEA/AIE-*2001 *The Fourteenth International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems*, June 4-7, 2001, Budapest in Monostori, L, Váncza, J., Ali Moonis (eds.) *Lecture Notes in Artificial Intelligence*, Springer, pp. 269-280, 2001.

[s8] **J. Valyon** and G. Horváth, „Reducing the complexity and network size of LS-SVM solutions”, POSTER: *NATO-ASI Conference*, Leuven, 2002.

[s9] **J. Valyon** and G. Horváth, „A generalized LS–SVM”, *SYSID'2003,* Rotterdam, pp. 827-832, 2003.

[s10] **J. Valyon** and G. Horváth, „A Sparse Least Squares Support Vector Machine Classifier”, *Proceedings of the International Joint Conference on Neural Networks IJCNN 2004*, Budapest, pp. 543-548, 2004.

[s11] G. Horváth**, J. Valyon**, Gy. Strausz, B. Pataki, L. Sragner, L. Lasztovicza, N. Székely, „Intelligent Advisory System for Screening Mammography”, *Proceedings of IMTC/04, 21th IEEE Instrumentation and Measurement Technology Conference*, Como, Italy, pp. 2071-2076, 2004.

[s12] **J. Valyon** and G. Horváth, „A Robust LS-SVM regression**”**, *Enformatica conference,* Prague, pp. 148-153, 2005.

[s13] **J. Valyon** and G. Horváth, „A Sparse Robust Model for a Linz-Donawitz Steel Converter”, *IEEE Instrumentation and Measurement Technology Conference*, Poland, May 1-3, 2007.

6.4. Local conferences

[s14] **J. Valyon** and G. Horváth, “Employing Support Vector Machines for Nonlinear System Modelling”, *Proceedings of the Mini-Symposium 2001*, pp. 34-35, 2001.

[s15] **J. Valyon** and G. Horváth, “A comparison of the SVM and LS–SVM regression, from the viewpoint of parameter selection”, *Proceedings of the 2002 Mini-Symposium*, pp. 18-19, 2002.

[s16] **J. Valyon** and G. Horváth, „A Weighted Generalised Least-Squares Support Vector Machine**”,** *Proceedings of the 10th Phd Mini-Symposium*, pp. 30-31., 2003.

[s17] **J. Valyon** and G. Horváth, „A Sparse Least Squares Support Vector Machine Classifier”, *Proceedings of the John von Neuman Phd Conference*, pp. 23-26, 2003.

[s18] **J. Valyon** and G. Horváth, „Controlling the complexity and network size of LS–SVM solutions”, *$(CS)^2$ - The Third Conference of PhD Students in Computer Science*, Szeged, Hungary, pp. 105, 2002. (best paper award).

7. REFERENCES

[1] R. Pintelon, I. Kollar, „On the frequency scaling in continuous-time modeling”,IEEE Transactions on Instrumentation and Measurement, 54: (1), pp. 318-321, 2005.

[2] J. Schoukens, R. Pintelon , T. Dobrowiecki and Y. Rolain, ”Identification of linear systems with nonlinear distortions”, Automatica 41, pp. 491-504, 2005.

[3] B. Pataki, „Neurális hálózatok alkalmazása dinamikus rendszerek modellezésében (Application of Neural Networks in the Modelling of Dynamic Systems)”, Doctoral dissertation, BME, Budapest, 1997. (in Hungarian)

[4] G. Horváth, „Neural Networks in Systems Identification”, (Chapter 4. in: S. Ablameyko, L. Goras, M. Gori and V. Piuri (Eds.) Neural Networks in Measurement Systems) NATO ASI, IOS Press, (Amsterdam, The Netherlands), pp. 43-78,2002.

[5] M. Altrichter, G. Horváth, B. Pataki, Gy. Strausz, G. Takács, J. Valyon, „Neurális hálózatok”, Panem, 2006.

[6] G. Horváth, B. Pataki, and Gy. Strausz, "Black box modeling of a complex industrial process", Proc. of the 1999 IEEE Conference and Workshop on Engineering of Computer Based Systems, Nashville, TN, USA, pp. 60-66, 1999.

[7] J. Van Gorp, "Nonlinear Identification with Neural Networks and Fuzzy Logic", Phd. Thesis, Vrije Universiteit Brussel, 2000.

[8] V. Vapnik, "The Nature of Statistical Learning Theory", New–York: Springer–Verlag., 1995.

[9] J. C. Platt. “Sequential Minimal Optimization: Fast Algorithm for Training Support Vector Machines”, Microsoft Research Technical Report MSR–TR–98–14, 1998.

[10] Yuh – Jye Lee and O. L. Mangasarian, “RSVM: Reduced Support Vector Machines”, Proceedings of the First SIAM International Conference on Data Mining, Chicago, 2001.

[11] J. A. K. Suykens, V. T. Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, “Least Squares Squares Support Vector Machines”, World Scientific, 2002.

[12] G. C. Cawley and N. L. C. Talbot, “Efficient formation of a basis in a kernel induced feature space”, In Proceedings of the European Symposium on Artificial Neural Networks (ESANN-2002), pages 1–6, Bruges, Belgium, April 24–26. 2002.