Paper

The Learning System by the Least Squares Support Vector Machine Method and its Application in Medicine

Paweł Szewczyk and Mikołaj Baszun

Faculty of Electronics and Information Technology, Warsaw University of Technology, Warsaw, Poland

Abstract—In the paper it has been presented the possibility of using the least squares support vector machine to the initial diagnosis of patients. In order to find some optimal parameters making the work of the algorithm more detailed, the following techniques have been used: K-fold Cross Validation, Grid-Search, Particle Swarm Optimization. The result of the classification has been checked by some labels assigned by an expert. The created system has been tested on the artificially made data and the data taken from the real database. The results of the computer simulations have been presented in two forms: numerical and graphic. All the algorithms have been implemented in the C# language

Keywords—classification, Grid-Search, Particle Swarm Optimization, patients diagnosis, Support Vector Machine.

1. Introduction

Recently has been observed efforts to remote patients diagnosis by use of teleinformatic services. It obeys automatic self diagnosis by proper software which is placed on Web servers accessable for patients [1], [2]. Also proper software could be used by the dedicated physicians for automatic preselection of the remote patients who need special attention of the physicians [3]. Such software need be continuously improved to obtain valuable help; software must be based on rules verified by the dedicated physicians.

The aim of the work was to create a system in the form of software that works with a database, enabling efficient classification of data and the initial diagnosis of patients with Least Squares Support Vector Machine (LS-SVM) classifier [4]. In order to find the optimal parameters which define work of the software several techniques have been used. The purpose of the preliminary data processing is the transformation of the input data set, as a result the new set will be obtained, for which the classification algorithm solves the problem with less error or in shorter time [5]. The first step in medical data preprocessing were:

- normalization – the transformation the data after which values of the attributes are in the range [min, max]; in this paper values min = 0, max = 1 have been adopted;
- standardization – the transformation of data, after which values of the attributes have an expected average value of zero and standard deviation equal to unity.

The use of standardization is safer and usually doesn't lead to bad consequences, as it may happen the case of normalization.

The result of the classification has been checked by some labels assigned by an expert. The created system has been tested on the artificially made data and the data taken from the real database. The results of the computer simulations have been presented in two forms: numerical and graphic.

2. Measures of Quality Assessment Classification

The basic problem that appears when we try to assess the ability of generalization of researched models, is the choice of a measure which will be used to estimate this ability [6]. In this research were used two: classification accuracy and confusion matrix.

Classification accuracy determines what part of all cases were correctly classified. It is expressed in percentage. When the accuracy is larger, then the classifier is more effective. In some applications, the distinct between incorrect classifications may have meaning. For example, in medicine pass a sick patient into the health group is much more dangerous than reverse situation. In these situations we may use the confusion matrix. It is the square matrix, where rows correspond to correct decision classes, and the columns refer to the decisions predicted by the classifier. In the case of LS-SVM algorithm, which is binary classifier, the confusion matrix can be written as shown in Table 1.

Table 1
Confusion matrix

Original classes	Predicted classes by the classifier	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The names used in Table 1 are inspired by the medical terminology, as follows:

- TP (true positive) – number of correctly classified examples from selected class,
- FN (false negative) – number of incorrectly classified examples from this class, negative decision when the example is in fact positive,

– TN (true negative) – number of examples which are not properly allocated to the selected class (correctly rejected),

– FP (false positive) – number of examples which are wrongly assigned to the selected class, when in fact they does not belong to (false alarm).

Using this kind of cases classification, particular attention should be paid to those examples, which are marked as FN. These examples are very important, because they mean not detection the disease, which can have bad consequences.

3. Least Squares Support Vector Machine Simulations

The advantages of the nonlinear SVM classifier are its great ability to solve the classification problems. J. Suykens in [4] proposed the method, which is the modification of the algorithm SVM V. Vapnik's [7], by modifying the cost function. This idea transformed the problem from solving the quadratic programming problem to solving a set of linear equations. This approach will simplify and shorten computation time.

3.1. Optimization of Model Parameters

In the case of the algorithm LS-SVM with radial kernel function [4], [7], optimized parameters are: γ, which is the weight at which the testing errors will be treated in relation to the separation margin and parameter σ, which corresponds to the width of the kernel function. It is not known in advance what combination of these two parameters will achieve the best result of classification. It is impossible to complete the search space of models, therefore the choice of optimal set of parameters is a very complex problem, and the way its solution is a key element of the classification system. In order to find the best values the following techniques were used: Grid-Search [8], K-fold Cross-Validation [5], Particle Swarm Optimization [9]–[15].

3.2. Used Technologies

To create an application development environment Microsoft Visual Studio 2010 has been used. As a programming language was chosen C# 3.0. The following libraries have been used:

– Windows Forms – implementation of the graphical user interface,

– ZedGraph – implementation of graph,

– ILNumerics – mathematical operations.

3.3. Characteristics of the Calculation Results

In order to test the proposed system, several data sets have been selected from UC Irvine Machine Learning Repository [16].

3.3.1. SPECT

A problem of diagnosis perfusion, based on data collected in Medical College of Ohio relies on diagnose of this disease based on 22 attributes [17], [18]. The database consists of 267 cases. All attributes take binary values. A specific case of cardiac perfusion may be classified to two classes: normal and abnormal. The division set to the classes was presented in the Table 2.

Table 2
Partition of the training and the test set – SPECT

Data	Train data		Test data	
Number of instances	80		187	
Perfusion	Normal	Abnormal	Normal	Abnormal
Number of instances	40	40	15	172

It may be noted that the dominant class in the test set are cases of incorrect perfusion – 91.9%. Moreover, in the process of learning classifier was used less examples than during testing. Therefore the subject of analysis it was how the algorithm can handle with a small number of training data, and how the cases will be diagnosed as correct perfusion, because such examples are only 8.1%. For this dataset there was no need for normalization and standardization because all attributes are binary.

Optimization with algorithm Grid-Search. As it has been studied in [13], the best method to find the optimal pair of parameters is changing them by exponential growth method. Using this technique, should be selected a pair of coefficients of γ and σ, for which the classification accuracy is the best.
In order to find the best model, it should be chosen the pair of points for which one side can get the highest accuracy, on the other cases, the best recognition of patients belong to the sick group. A compromise, between perhaps sometimes inconsistent conditions, is necessary.
Making an analysis of the obtained curves the point of coordinates: $\gamma = 2^{52.2}$ and $\sigma = 2^{-1.1}$ has been chosen. In the case someone can observe a slightly lower classification accuracy, but it is maximized the value of parameter true positive. For these data the aim is the best diagnosis of the sick patients. The measure of this recognition is the TP factor, so it should endeavor to a situation when

Table 3
Results of classification on a test set with Grid-Search

Parameter	Value
$\mathrm{Log}_2\gamma$	52.2
$\mathrm{Log}_2\sigma$	–1.1
Classification accuracy [%]	90.9
TP	162
FP	7
TN	8
FN	10

this coefficient has as the greatest value as possible. For this value will be followed a classification with LS-SVM algorithm on a test set.
In Table 3 the results of the classification with LS-SVM algorithm on a test set are presented – data distribution has been shown in Table 2. It is worth noting that only 10 of 172 cases have not been diagnosed as the sick persons. On the other side, the 7 from 15 people have been wrongly diagnosed as sick, although they should be included into the healthy group.

Optimization with algorithm Particle Swarm Optimization. After a large number of simulations with varying parameters of PSO algorithm, the results have been obtained, some of which are presented in the Table 4. It has been studying, how the number of partitions and the number of iterations affect on the obtained results.

Table 4
Results of classification on a test set with PSO

Number of partitions	50	50	100	100
Number of iterations	50	100	50	100
$\text{Log}_2\gamma$	48.8	48.5	24.8	50.4
$\text{Log}_2\sigma$	1.1	–1.9	–1.9	–0.6
Classification accuracy [%]	79.14	89.30	89.83	93.58
TP	138	159	160	168
FP	5	7	7	8
TN	10	8	8	7
FN	34	13	12	4

In the Table 4 the results of the classification with LS-SVM algorithm on a test set have been presented – data distribution has been shown in Table 2. The good level of accuracy has been achieved. Only 4 of 172 cases were not diagnosed as sick. In turn, 8 of 15 people were wrongly diagnosed as sick, although they should be included into the healthy group. The result for the patients from the sick group is 97.6%, whereas for the patients from the healthy group is 46.6%. Making an analysis of the above results, it can conclude, that the algorithm coped very well with the diagnosis of the cases from the sick group. It is evident that the number of particles and iterations allows to get better results. With a small number of particles the algorithm is stagnant and the particles can't escape from local minima. The values of the parameters c_1 and c_2 are not as important for the convergence of the algorithm. The experimental results indicate that it is better to set the value of parameter w to the large one in order to promote a global exploration of the space, and gradually decrease it to obtain more improved solutions. The initial value was set to 0.9 and then was reduced, in each step, to the value 0.4 [14].

3.3.2. Breast Cancer

A problem of diagnosis breast cancer, based on data collected in University Hospital in Madison (Wisconsin) relies on a diagnose of this disease based on 30 attributes [19], [20]. The database consists of 569 cases. It has no missing attribute values. A specific case of cardiac perfusion may be classified to two classes: malignant and benign. The division set of the class has been presented in Table 5.

Table 5
Percentage distribution of classes in the dataset – breast cancer

Class attribute	Number of cases
–1 (malignant cancer)	212 (37.2%)
1 (benign cancer)	357 (62.8%)

In the Table 5 the characteristics of the data set has been presented. The dominant class in these input set are cases of benign cancer – 62.8%. Because the data have not been pre-divided into training and testing datasets, the effect of the percentage partition of the data on the classification accuracy will be investigated. Each time a training set will be selected at random and will contain from 10% to 90% of all data. For each step it had been carried out 10 drawings of the training set, and then the average classification accuracy has been calculated.

Optimization with algorithm particle swarm optimization. In the simulations the following parameters were adopted in the PSO algorithm: $c_1 = c_2 = 0.5$, w-according to the method described above, number of partitions = 100, number of iterations = 100. The obtained results have been shown graphically in Fig. 1. It presents the averaged classification results using several methods of the preliminary data processing.

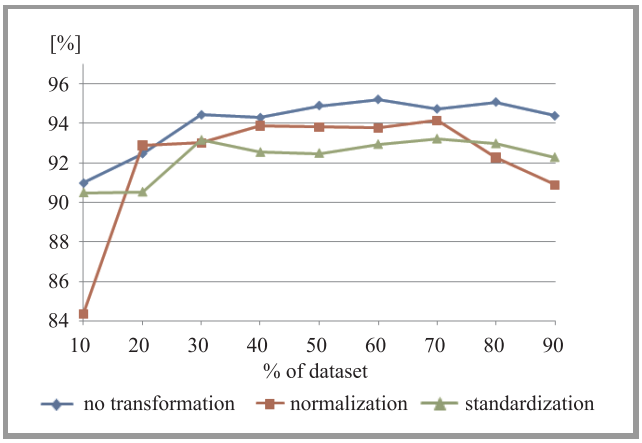


Fig. 1. Comparison of average classification accuracy for three methods of data preprocessing when changes the training set size – breast cancer.

In Fig. 1 it can be seen that the best classification accuracy was obtained when it was not used any preprocessing method. It is evident that when the training set is less than 10%–20%, the obtained results are worse. This is due to the fact the border decision boundary in LS-SVM algorithm, what depends largely on the representativeness of training data. If they are not enough, it can't achieve a satisfactory level of generalization of the model.

3.3.3. Heart Disease

A problem of diagnosis heart disease, based on data collected in Hungarian Institute of Cardiology, University Hospital in Zurich, University Hospital in Basel, V. A. Medical Center in Long Beach, Cleveland Clinic Foundation, relies on a diagnose of this disease based on 13 attributes [21]–[24]. The database consists of 270 cases. It has no missing attribute values. Each specific case may be classified to two classes: healthy and sick. The division set to the class was presented in Table 6.

Table 6
Percentage distribution of classes in the dataset – heart disease

Class attribute	Number of cases
–1 (sick)	120 (44.4%)
1 (healthy)	150 (56.6%)

In Table 6 the characteristics of the data set has been presented. The dominant class in this input set are cases of healthy patients (55.6%). Because the data have not been predivided into training and testing datasets, investigated the effect of the percentage partition of the data on the classification accuracy. Each time a training set will be selected at random and will contain from 20% to 90% of all data. For each step was carried out 10 drawings the training set, and then the average classification accuracy have been calculated.

Optimization with algorithm particle swarm optimization. In the simulations the following parameters were adopted in the PSO algorithm: $c_1 = c_2 = 0.5$, w-according to the method described above, number of partitions = 100, number of iterations = 100. The obtained results have been shown graphically in Fig. 2. It presents averaged classification results using several methods of the preliminary data processing.

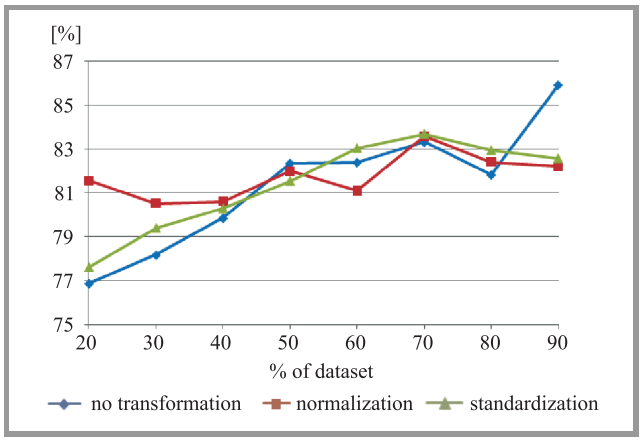


Fig. 2. Comparison of average classification accuracy for three methods of data preprocessing when changing the training set size – heart disease.

In Fig. 2 it can be seen that the best classification accuracy was obtained used the pre-processing method, which are similar. The results accuracy are lower than it was in the case of the breast cancer dataset. This may be due to fewer examples available in the database, and hence smaller number of examples used in the learning process.

4. Conclusions

The aim of the presented work was the adaptation of algorithmic techniques LS-SVM to their most efficient using in the classifying medical data from the patients. The idea here it is primarily to allow the software classification can be reliably put the presumptive diagnosis. The work of this software cannot substitute a real expert – a doctor, but to support his work by patients' selection. The process of classification of patients, based on medical data, was in the work carried out in such a way, that it analyzed data from actual patients, who were already known, what is the correct medical diagnosis. This was possible due to the using of databases available on the Internet, put there by reputable clinics.

The process of adapting the algorithm LS-SVM consisted primarily a very time consuming repetition of the calculations for other sets of parameter values which define the work of computational process, and then the choice of the most favorable version from the point of view of accuracy assessment of quantitative decision making processes results. It can't be predicted analytically for SVM techniques, as well as for other groups of algorithms based on the philosophy of artificial neural networks.

As a result, it seems that the group received the application software suitable for using in a practical analysis of data from a database of medical patients. It was found that the analysis of different groups of medical data classification software by the LS-SVM method has to be differentiated. This was demonstrated by analyzing the sample data on several key areas of treatment.

It was also studied the effect of normalization and standardization of data on the final effect of the allocation of patients. It can't determine what method of data preprocessing is better. It depends on the specific data. And so, in the case of breast cancer database, it was found that by using the normalization and standardization worse results than without preprocessing were achieved. Differences in values are on a level of a few percent. And in the case of heart disease when the training set was smaller, better results were obtained, by using normalization. When the number of examples increased, the results obtained were very close to each other.

Acknowledgement

This work was partly supported by founds on science in 2007–2010 as Ordered Research Project of Polish Ministry of Science and Higher Educations.

References

[1] M. Baszun and B. Czejdo., "An interactive medical knowledge assistant", in *Visioning and Engineering the Knowledge Society*, Berlin: Springer, 2009.

[2] M. Baszun, "Real time medical advising in cyberspace and its security aspects", in *Proc. VI Int. Conf. Cyberspace 2009*, Brno, Czech Republik, 2009.

[3] M. Baszun and B. Czejdo, "Remote patient monitoring system and a medical social network", *Int. J. Social Humanistic Comput.*, vol. 1, no. 3, pp. 273–281, 2010.

[4] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*, World Scientific Publishing Company, 2002.

[5] N. Jankowski, "Ontogeniczne sieci neuronowe w zastosowaniu do klasyfikacji danych medycznych", Praca doktorska, Katedra Metod Komputerowych Uniwersytetu Mikołaja Kopernika, Toruń, 1999.

[6] P. Cichosz, *Systemy Uczące się*. Warszawa: WNT, 2000 (in Polish).

[7] C. Cortes and V. Vapnik, "Support-vector network", *Machine Learning*, vol. 20, pp. 273–297, 1995.

[8] C.-W. Hsu, C.-C. Chung, C.-J. Lin, *A Practical Guide to Support Vector Classification*, National Taivan University, March 13, 2010 [Online]. Available: www.csie.ntu.edu.tw/∼cjlin

[9] S. Das, A. Abraham, and A. Konar, *Particle Swarm Optimization and Differential Evolution Algorithms: Technical Analysis, Applications and Hybridization Perspectives*. Berlin: Springer, 2008.

[10] J. Keneddy and R. C. Eberhart, "Particle swarm optimization", in *Proc. IEEE Int. Conf. Neural Netw.*, Piscataway, New York, pp. 1942–1948, 1995.

[11] M.-Z. Lu, C. L. Philip Chen, J.-B. Huo, "Optimization of combined kernel function for SVM by particle swarm optimization", in *Proc. Eighth Int. Conf. Machine Learning Cybernet.*, Baoding, China, 2009, pp. 1160–1166.

[12] M. G. H. Omran, "Particle swarm optimization methods for pattern recognition and image processing", Ph.D. thesis, University of Pretoria, 2004.

[13] C. Sun and D. Gong, "Support Vector Machines with PSO Algorithm for Short-Term Load Forecasting", in *Proc. IEEE Int. Conf. Netw., Sensing Contr. ICNSC '06*, Ft. Lauderdale, USA, 2006, pp. 676–680, 2006.

[14] Q.-Z. Yao, J.e Cai, J.-L. Zhang, "Simultaneous feature selection and LS-SVM parameters optimization algorithm based on PSO", in *Proc. World Congr. Comput. Sci. Informa. Engin. CSIE 2009*, Los Angeles, USA, 2009, pp. 723–727.

[15] Y.g-J. Zhai, H.-L. Li, Q. Zhou, "Research on SVM algorithm with particle swarm optimization", in *Proc. 11th Joint Conf. Inform. Sci. JCIS 2008*, Shenzhen, China, 2008.

[16] *UC Irvine Machine Learning Repository*, Center for Machine Learning and Intelligent Systems, University of California, USA [Online]. Available: http://archive.ics.uci.edu/ml/

[17] L. A. Kurgan, K. J. Cios, R. Tadeusiewicz, M. Ogiela, and L. Goodenday, "Knowledge discovery approach to automated cardiac SPECT diagnosis", *Artificial Intelligence in Medicine*, vol. 23, no. 2, pp. 149–169, 2001.

[18] A. Płachcińska and J. Kuśmierek, *Techniki Obrazowania Serca w Medycynie Nuklearnej*, Zakład Medycyny Nuklearnej Akademii Medycznej w Łodzi, 2001 (in Polish).

[19] K. Polat and S. Güneş, "Breast cancer diagnosis using Least Square Support Vector Machine", *Digit. Sig. Proces.*, vol. 17, pp. 694–701, 2007.

[20] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis", in *Proc. Int. Symp. Electron. Imag.: Sci. Technol.*, San Jose, USA, 1993, vol. 1905, pp. 861–870.

[21] N. Allahverdi and H. Kahramanli, "Extracting rules from neural networks using artificial immune systems", in *Proc. 2nd Int. Conf. Problems of Cybernet. Inform.*, Baku, Azerbaijan, 2008.

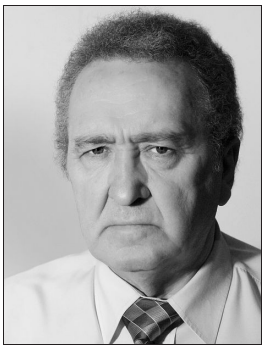
[22] S. Bhatia and P. Prakash, "SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features", in *Proc. World Congr. Engin. Comp. Sci. WCECS'08*, San Francisco, USA, 2008.

[23] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques", *Int. J. Comput. Sci. Network Secur.*, vol. 8, no. 8, 2008.

[24] D. W. Vance, *An All-Attributes Approach to Supervised Learning*, University of Cincinnati, 2006.

Paweł Szewczyk received the M.Sc. degree in Electronics and Computer Engineering from Warsaw University of Technology in 2011. His research interests focus on neural networks, support vector machines, especially in classification and data mining.

E-mail: pawelszewczyk1@gmail.com
Faculty of Electronics and Information Technology
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Mikołaj Baszun is an Assistant Professor in the Faculty of Electronics and Information Technology, Warsaw University of Technology. His research focuses on electronic microsystems, computer engineering, artificial intelligence technology, medical informatics, and Web services. He has published more than 50 publications in these areas.

E-mail: mbaszun@elka.pw.edu.pl
Faculty of Electronics and Information Technology
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland