Application of support vector machine for prediction and classification

V A Kramar¹, V V Alchakov¹, V R Dushko¹, T V Kramar²

¹ Sevastopol State University, 33, Universitetskaya St., Sevastopol, 299053, Russia
² Crimean Federal University V.I. Vernadskogo, 4, Vernadskogo Ave., Simferopol, 298400, Russia

E-mail: kramarv@mail.ru

Abstract. The article deals with the technology of vector machines for solving prediction and classification problems. The specialized software that used LibSVMsharp library is represented.

1. Introduction
The development of computing power of personal computers has opened new opportunities in solving problems associated with the processing of time series. In particular, prediction and classification problems can be solved efficiently [1-5].

The solution of the prediction problem is reduced to calculating the approximate or predicted value of a function:

\[
y = f(x_1, x_2, \ldots, x_n)
\]

that depends on a set of parameters \(x_i, i \in [1 \ldots n]\) available for measurement or calculation based on a set of measured data.

Certain values of the parameters are specified in technical and economic systems. In turn, the problem of classifying the individual events or events is characterized by a set of parameters, reduced to the problem that deal with two given classes. For example, the conclusion about the presence or absence of the disease is based on analysis results. In addition, a classifying function of form (1), depending on a set of parameters or attributes \(x_i, i \in [1 \ldots n]\), is also introduced.

To solve this type of problem, the following approaches could be used:
- neural networks;
- support vector machines;
- regression analysis;
- associative rules, etc.

In next section, only one from the mentioned methods will be used. This method is called a support vector machine (SVM) method.

2. Support Vector Machine method
In its present form, the method was designed in 1995 by V.N. Vapnik who worked in «AT&T Bell Laboratories» corporation and was well known as a scientist in the field of machine learning [6]. Early, the method was positioned as an algorithm based on training with the teacher to solve the
classification problem, but later it was also used to perform regression analysis. The main idea of the
method is to construct an optimal separating hyperplane in the space of high-dimensional features. Optimality is understood in the sense of minimizing the upper bounds for the probability of a
generalization error.

In this way, the problem of binary classification should be considered. There are only two classes
of objects and it is necessary to determine whether the object belongs to a particular class according
to a given set of parameters. Each classification object \( X \) is a vector in the \( n \)-dimensional space
\([x_1, \ldots, x_n] \in \mathbb{R}^n\), while each coordinate of this vector is a certain attribute of the object that should be
set or defined. There is the direct dependency between the coordinate value and the expression of the
sign in the object.

A certain set, which consists pair combinations, is considered:
\[ XY = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_m, Y_m)\}, \]
where \( Y_i = y_i \in \{-1; 1\} \). This set can be used for training.

The task is reduced to finding a rule, according to which any arbitrary classification object \( X \)
specified by a set of parameter values \([x_1, \ldots, x_n]\) can be assigned to one of the classes \((1 \text{ or } -1)\). Such
rule can be described using the following expression:
\[
F(X) = \sum_{i=1}^{n} \omega_i x_i + b = W^T X + b
\]
\[
Y(X) = \begin{cases} 
1, & \text{if } F(X) > 0 \\
-1, & \text{if } F(X) < 0 
\end{cases},
\]
where \( \omega_i \) – weights coefficients, \( b \) – shift parameter.

From the geometric point of view, the linear classifier corresponds to some separating hyperplane
\( F(X) = 0 \) in space \( \mathbb{R}^n \), while the object belongs to the first class if it is located on the positive side of
the hyperplane, and otherwise – to the second class, as shown in Figure 1.

![Figure 1. The separating hyperplane of a linear classifier](image)

The initial data for the prediction problem with the support vector method are the same as for the
classification problem. It is required to define function \( F(X) \) that can be used for the best
interpolation for all elements in the sample set:
\[ XY = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_m, Y_m)\}, \]
where \( Y_i = y_i \in \mathbb{R} \).

Then function \( F(X) \) is used to obtain approximate value \( \hat{Y} \) for arbitrary parameter vector
The problem of finding the parameters of function  \( F(X) \) is formulated as a minimization of the functional:

\[
\frac{1}{2} W^T W + C \sum_{k=1}^{n} (\xi_k + \xi_k^*) \rightarrow \min
\]

with limitations:

\[
\begin{align*}
Y_k - (W^T \phi(X_k) + b) & \leq \varepsilon + \xi_k, \\
(W^T \phi(X_k) + b) - Y_k & \leq \varepsilon + \xi_k^*, \\
Y_k (W^T \phi(X_k) + b) - Y_k & \geq 1 - \xi_k^*, \\
\xi_k, \xi_k^* & \geq 0
\end{align*}
\]

where positive constant \( C \) is a regularization parameter, and the larger its value is, the greater the error penalty is. \( \xi_k, \xi_k^* \) are variables that determine the upper and lower boundaries of the learning error, respectively. \( \varepsilon \) is the loss function, which is often being an piecewise linear function.

While the support vector machine method is used, the problem of finding functional dependency is resolved:

\[
F(X) = \sum_{k=1}^{n} (\alpha_k - \alpha_k^*) K(X_k, X) + b
\]

There the dual variables satisfy condition \( 0 < \alpha_k, \alpha_k^* \leq C \), and so-called symmetric nuclear function \( K(\cdot, \cdot) \) satisfies the Mercer’s conditions [7]. If non-zero values \( \alpha_k \) match the data elements, they are named the reference vectors.

In this case, the determination of parameter value \( C \), as well as the determination of the type of nuclear function \( K(\cdot, \cdot) \), largely determines how accurately the regression model will be obtained to predict the function value by the use of a set of specified parameters.

### 3. The Software realization of the SVM method

To solve problems (2) – (5), there are many programming libraries. However the most common library is libsvm (Library for Support Vector Machines) that is developed by the scientists from the National Taiwan University: Chih-Chung Chang and Chih-Jen Lin [8,9]. This library can be used both in software products for scientific and in engineering calculations such as Matlab and R; also it can be used for the development of native software in high-level languages C ++ or Java. However, there is no possibility of its direct connection to the software solutions that are based on Microsoft .NET technology.

Nevertheless, this problem can be solved using the open library LibSVMsharp [10] used as a program bridge between libsvm and .NET level, and this one can be used for the SVM method realization in C# based applications.

In this research, a software product that is designed to process experimental data in order to solve the problem of classification and forecasting was developed. Thus, the software product can be used in various fields of knowledge such as medicine, design of control systems, design of technical systems operating in extreme conditions, military science, etc.

The main application window is represented in Figure 2. To develop friendly base GUI, the DevExpress controls library was used as well. Application was compiled in Visual Studio 2015 and require .NET Framework 4.5.2 installed on an end-user machine. For data storage SQLite, the local database was used. No other libraries or frameworks should be installed for correct application functionality.
A working space of the main window is divided into three parts (to separate graphic and text information).

The graphical part displays two time series that correspond to the true data and estimates and is obtained using SVM, as well as a graph of the calculation error. The text part is represented in the form of tabulated function values and estimations that are obtained with the help of SVM, and calculation errors. Statistical processing of the obtained results is also carried out, specifically calculation of the root-mean-square error of calculations, dispersion, maximum, minimum and average error values.

For parameter tuning a modal settings window is used. Setting options allows one to add, delete, edit, change settings, select the type of SVM, specify the type of the kernel of the machine. These sets of parameters will be used during the training and cross-validation of the model, in order to determine the optimum values from the point of view of minimizing calculation errors. All intermediate calculations and the received SVM models are saved in separate files, which allows performing the subsequent analysis of the results and working out recommendations for the optimal adjustment of SVM and improving the quality of the received forecasts.

The tuning set could be saved in the xml format in the application folder and this set (or any other) can be used whenever a user wants to apply this set for some data set. There is no some universal set that could be used for any data set. So it is very important to find the optimal SVM parameters set that will be applied to the current data to receive best statistical results.

All data can be summarized to one report in the pdf format, which is automatically generated upon completion of the training and testing of the model.
4. Conclusion
The software implementation of the support vectors machine method for solving the prediction and classification problem is performed. The further direction of the research is associated with automation of the selection of optimal values for SVM parameters for improving the quality of the received forecasts.

References
[1] Hsu C, Lin C 2002 A comparison of methods for multiclass support vector machinesIEEE Transactions on Neural Networks13(2) 415–425
[2] Kramar V, Dushko V, Rodkina A, Alchakov V 2016 Neural network modeling the potential of steels with different chrome contentProceedings of the 26th DAAAM International Symposium 0641-0647
[3] Sudheer C, Sohani S, Kumar D 2014 A support vector machine-firefly algorithm based forecasting model to determine malaria transmissionNeurocomputing 129 279–288
[4] Chen H, Yang B, Wang S, 2014 Towards an optimal support vector machine classifier using a parallel particle swarm optimization strategyApplied Mathematics and Computation 239 180–197
[5] Wu C, Tzeng G, Goo Y, Fang W 2007 A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcyExpert Systems with Applications 32(2) 397–408
[6] Vapnik V N 2000The nature of statistical learning theory (information science and statistics). 2nd ed. (N.Y.: Springer-Verlag)
[7] Sheremet A I, Sadovoy A V 2014 Regression analysis in interconnected electromechanical systems based on the support vector methodInformation and control systems 1(68)48–56
[8] Chang C C, Lin C J 2013A Library for Support Vector Machines. Retrieved from https://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf
[9] HsuCW, Chang C C, Lin C J 2016 A Practical Guide to Support Vector Classification. Retrieved from https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
[10] Erhan C 2014 C# wrapper for Support Vector MachinesRetrieved from https://github.com/ccerhan/LibSVMsharp