Support vector machines-type procedure through kriging

M. A. Matin*, K. V. Mardia, & C. C. Taylor

University of Leeds

Good science is the ability to look at things in a new way and achieve an understanding that you didn't have before ... It is opening windows on the world ... you perceive a little tiny glimpse of the way the Universe hangs together, which is a wonderful feeling. Hans Kornberg

Support vector machines are used as a universal constructive learning procedure based on the statistical learning theory developed by V. Vapnik (Vapnik, 1995). Recently several research groups have shown excellent performance of SVMs on different problems of classification and regression. Much of the works done in SVM stems from the disciplines of Computer Science, Infromation Science, Machine Learning, Engeneering, Computer Engeneering, etc rather Statistics or even Mathematics. Statisitcal works are very limited. The present work develops SVM-type procedure within a kriging framework for carrying out classification and prediction. This study is motivated by microarray gene expression data.

The genome is the total genetic information that an organism possesses. Genes are made from DNA, which contains the complete genetic information that defines the structure and function of an organism. DNA stores information in the form of the base nucleotide sequence, which is a string of 4 letters (Adenine, Cytosine, Guanine, and Thymine). Genes are segments of DNA that contain the "recipe" to make proteins and proteins are the crucial molecules that do most of a cell's work. Indeed, the "central dogma of molecular biology" states that, in a cell, information flows from the nuclear DNA to RNA to protein synthesis. Hence proteins are formed using the genetic code of DNA, and a protein sequence can be represented by a string of 20 letters, each standing for an amino acid. (see for example Lesk 2002, Mardia, et. al. 2003). A microarray measures the number of copies of messenger RNA (mRNA) in a given sample of cells. The microarray is appealing because of its ability to produce data in a high-throughput fashion. Gene expression data on p genes for n tumor samples is summarized by an $n \times p$ matrix, so x_{ij} is the measuement of the expression level of the jth gene for the ith sample where $j = 1, \ldots, p$.

A support vector machine finds a non-linear decision function in the input space by mapping the data into a higher dimensional feature space and separating it there by means of a maximum margin hyperplane. The computational complexity of the classification operation does not depend on the dimensionality of the feature space, which can even be infinite. Overfitting is avoided by controlling the margin. The separating hyperplane is represented sparsely as a linear combination of points. The system automatically identifies a subset of informative points and uses them to represent the solution. Finally, the training algorithm solves a simple convex optimization problem. All these features make SVM an attractive classification system. See Brown, et al. (2000), Mallik, et al. (2003), Wahba, et al. (2002), Cortes and Vapnik (1995).

Mallik et al. (2003) have propsed a reproducing kernel Hilbert space based classification method for microarray data. It is shown that these models in a Bayesian hierarchical setup with priors over the shrinkage (smoothing) parameters performed better than the other popular classification methods. Also, multiple shrinkage models always appear to be superior to single parameter shrinkage models. With multiple shrinkage parameters, the regular Bayesian SVM model emerges as the winner in all the examples with the Complete SVM finishing a close second all the time. However, the Complete SVM provides a more formal probabilistic motivation for the use of SVM's, and are more satisfactory from a Bayesian angle perspective.

Kriging is the name of a technique developed by Matheron in the early 1960s for mining applications. See for example Ripley (1981), Cressie (1993), Chilie's and Delfiner (1991). The kriging model postulates a combination of a known function and departures of the form $y(x) = f(x) + W(x)$ where $y(x)$ is the unknown function of interest, $f(x)$ is a known polynomial function, which is often taken as a constant, and $W(x)$ is called the correlation function and is a realization of stochastic process with mean 0 and variance σ^2, and nonzero covariance. Flexibity in kriging is achieved through a variety of spatial correlation functions, but the Gaussian correlation function is most frequently used. In kriging we need to use covariance matrix which is identical to kernel function or at least act as a kernel in kriging.

Interest is in modelling the data, i.e. estimating distributional parameters, and then to predict the phenomenon under study at unobserved sites within the corresponding sampling domain. The method of universal kriging for spatial prediction was introduced to cover the problem of spatial trend effects. This is done by incorporating linear trend models, e.g., polynomial functions of the spatial coordinates.

In order to classify data into one of two groups with respective class probability the logistic regression model may be used with some threshold (say 50%) to decide to predict for that class case. However, the explanatory variables X's may be covariates: could be (spatial) location but fixed and it could be possible to observe the response variable Y for the explanatory variables X's. So possible structure/model of spatial location (fixed) could be investigated using the covariance structure of X's based on distance function. Therafter, logistic regression may be used to predict the responses.

Our motive is to the use the kriging methodolgy of classification and prediction to achieve the same goal as of SVM perhaps with a greater computational and cost benifits. In this regard the microarrary data can be seen as of the n pairs $(y_i, x_i), i = 1, 2, \ldots, n$, where each pair is a location x_i, and the observed data value y_i measured at this location. It is assumed here that x_i is a p-dimensional vector. Therefore, a typical and minimal data structure would consist of a vector and a matrix with p columns. That is the data object is a list of two components : (i) "coordinates" which is an $n \times p$ matrix and (ii) "data" which is an n-dimensional vector. Just like we may mimic the idea of microarray gene expression data. In this regard we will exploit the logistic regression model as classifier and the Gaussian (isotropic) process.

The inner-product kernel performing the nonlinear mapping into feature space is

$$K(x, x_i) = K(x_i, x) = \varphi(x)^T \varphi(x_i)$$

And the total optimization problem is

$$Minimize\ Q(\alpha) = \sum_{i=1}^n \alpha_i - 1/2 \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

subject to some constraints. The only requirement on the kernel $K(x, x_i)$ is to satisfy Mercer's theorem.

And in the trend surface with degree zero. Suppose that $f(x)$ is known, so that we can work with $W(x) = f(x) - y(x)$. We will consider only linear predictors

$$\widehat{W}(\mathbf{x}) = \sum \lambda_\mathbf{i} \mathbf{W}(\mathbf{x_i}) = \lambda^\mathbf{T} \mathbf{W_N}$$

and choose that which minimizes

$$\begin{aligned} E(\mathbf{W}(\mathbf{x}) - \widehat{\mathbf{W}}(\mathbf{x}))^\mathbf{2} &= \text{var}(\mathbf{W}(\mathbf{x})) - \mathbf{2} \sum \lambda_\mathbf{i} \mathbf{C}(\mathbf{x_i}, \mathbf{x}) + \sum \lambda_\mathbf{i} \lambda_\mathbf{i} \mathbf{C}(\mathbf{x_i}, \mathbf{x_j}) \\ &= \sigma^2(\mathbf{x}) - \mathbf{2} \lambda^\mathbf{T} \mathbf{k}(\mathbf{x}) + \lambda^\mathbf{T} \mathbf{K} \lambda. \end{aligned}$$

where $K_{ij} = C(\mathbf{x}_i, \mathbf{x}_j), k(\mathbf{x}) = (C(\mathbf{x}, \mathbf{x})$, a column vector, and of course, λ depends on $\mathbf{x}$.

The above equations make clear the relationship between kriging and SVM. The dual problems for the static regression without bias term are closely related to Gaussian process (MacKay 1992), regularization networks (Poggio and Girosi 1990) and kriging (Cressie 1993), while LS-SVMs rather take an optimization approach with primal-dual formulation which have been exploited towards large scale problems and in developing robust versions.

Under the above perspective, we have SVM-type procedure within the framework of kriging for the classification and prediction. Some illustrative examples have been worked out to give insight into the new type of statistical SVM.

References

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C., Furey, T. S., Ares, JrM., and Haussler, D. (2000): Knowledge-based analysis of microarray gene expression data using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 262-267.

Chile's, J. P. and Delfiner, P. (1991). *Geostatistics*. New York: Wiley.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**, 273-279.

Cressie, N. A. C. (1991). *Statistics for Spatial Data*. New York: John Wiley and Sons.

Kornberg, H. (). Journal of Irreducible Results.

Lesk, A. M. (2002). *Introduction to Bioinfromatics*. Oxford, Oxford University Press.

MacKay, D. J. C. (1992). The evidence framework applied to classification networks. *Neural Computation*, **4**, 698-714.

Mallik, B., Ghosh, D., and Ghosh, M. (2003). Bayesian clssification of tumor using gene expression data. *Proceedings in Stochastic Geometry, Biological Structure and Images*, 107-110. Edited by R.G. Aykroyd, K.V. Mardia and M.J. Langdon. Leeds University Press.

Mardia, K. V., Taylor, C. C., Westhead, D. R. (2003). Structural bioinformatics revisited. *Proceedings in Stochastic Geometry, Biological Structure and Images*, 11-18. Edited by R.G. Aykroyd, K.V. Mardia and M.J. Langdon. Leeds University Press.

Matheron, G. (1960). The intrinsic random functions and their applications. *Advances in Applied Probability* **5**, 115-133.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Procceedings of the IEEE*, 1481-1497.

Ripley (1981). *Spatial Statistics*. New York: John Wiley and Sons.

Vapnik (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995 (2nd ed. in 2000).

Wahba, G., Lin, Y., Lee, Y., and Zhang, H. (2002). Optimal properties and adptive tuning of standard and nonstandard support vector machines. *Technical Report*, University of Wisconsin.