An Introduction to the Application of Support Vector Classification

Miguel Cuba, Enrique Gallardo and Oy Leuangthong

Sequential Indicator Simulation (SIS) and Truncated Gaussian Simulation (TGS) are used for building categorical models based on point data. The implementation of both requires the modeling of a variogram model which is a subjective and sometimes a time consuming task. This paper shows the implementation of a Support Vector Machine (SVM) for building categorical models, which does not require fitting any variogram model from the conditioning data. The advantages and disadvantages on the use of them, as well as two case studies are presented and discussed in the present document.

Introduction

The support vector algorithm (Boser, Guyon, & Vapnik, 1992) was initially developed for solving classification problems. Soon after, it was extended to deal with regression problems (Muller, Smola, Ratsch, Scholkopf, Kohlmorgen, & Vapnik, 1997). The SVC algorithm seeks for the weight parameters **w** and the bias term *b* of a decision boundary (a hyperplane) of the form:

$$\mathbf{w}^{\mathrm{T}}\mathbf{u} + b = 0 \tag{1}$$

The boundary will separate the categories given on the observed data with a maximum margin as illustrated in Figure 1. The decision boundary will classify the binary category *s* at unsampled locations **u** according to the rule:

$$s(\mathbf{u}) = \begin{cases} s_1 & if \quad \mathbf{w}^{\mathrm{T}}\mathbf{u} + b > 0 \\ s_2 & if \quad \mathbf{w}^{\mathrm{T}}\mathbf{u} + b < 0 \end{cases} \tag{2}$$

SVC has the following steps: (1) preprocessing of data, (2) SVC training, and (3) SVC testing.

Preprocessing of data

Let us consider a dataset has two categories, that is, s_1 and s_2. To perform SVC the set is coded as:

$$i(\mathbf{u}_\alpha) = \begin{cases} 1 & if \ s(\mathbf{u}_\alpha) = s_1 \\ -1 & \text{if } s(\mathbf{u}_\alpha) = s_2 \end{cases} \tag{3}$$

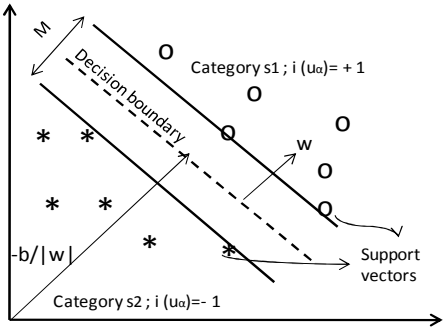

Figure 1: SVC concepts: codification (±1), margin (*M*), weights (w) and support vectors.

SVC training

Finding the weighting parameters **w** and the bias term *b* of the decision boundary (1) using the observed data is referred to as training the SVC. In machine learning jargon the observed data is called the training set. The percentage of observed data misclassified by the decision boundary is called training error or empirical error. The complement of the empirical error to add to unity is called in this thesis *empirical accuracy*.

The boundary (1) is determined to maximize the margin of separation between the categories s_1 and s_2 (see Figure 1). If the data is linearly separable, the optimization problem is expressed as:

$$\underset{\mathbf{w},b}{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2$$
$$subject\ to \quad i(\mathbf{u}_\alpha)[\mathbf{w}^{\mathrm{T}}\mathbf{u}_\alpha + b] \geq 1 \quad ; \quad \alpha = 1,\ldots,n \tag{4}$$

where ||**w**|| represents the Euclidean norm of the vector **w**. This nonlinear optimization problem with inequality constrains is solved using the Lagrange formalism and leads to the following results for **w** and *b*:

$$\mathbf{w} = \sum_{\alpha=1}^{n} \eta_\alpha i(\mathbf{u}_\alpha)\mathbf{u}_\alpha \tag{5}$$

$$b = \frac{1}{Nsv}\left(\sum_{\alpha=1}^{Nsv}\left(\frac{1}{i(\mathbf{u}_\alpha)} - \mathbf{u}_\alpha^{\mathrm{T}}\mathbf{w}\right)\right) \quad ; \quad \alpha = 1,\ldots,Nsv \tag{6}$$

where η_α are Lagrange multipliers and *Nsv* is the numbers of support vectors; that is, training data whose η_α are not zero. Substituting (5) into (1) the boundary becomes:

$$\sum_{\alpha=1}^{n} i(\mathbf{u}_\alpha)\eta_\alpha \mathbf{u}^{\mathrm{T}}\mathbf{u}_\alpha + b = 0 \tag{7}$$

An overlap of the categories may indicate that a plane that separates them does not exist. To deal with this case, the linear SVC was adapted (Cortes, 1995), (Cortes & Vapnik, Support Vector Networks, 1995) by the introduction of slack variables ξ_α (α = 1,…,*n*) in the optimization problem. The slack variables ξ_α relax the constraints in (4), therefore, some classification errors are permitted but at a certain cost. Now, the optimization problem is:

$$\underset{\mathbf{w},b,\xi}{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + P\sum_{\alpha=1}^{n}\xi_\alpha$$
$$subject\ to \quad i(\mathbf{u}_\alpha)[\mathbf{w}^{\mathrm{T}}\mathbf{u}_\alpha + b] \geq 1 - \xi_\alpha \tag{8}$$
$$\xi_\alpha \geq 0 \quad ; \quad \alpha = 1,\ldots,n$$

Here, *P* is a user-defined penalty parameter. The optimization problem has the same solution shown in (5) , (6) and (7), the only difference is the bounds of the multipliers η_α that appear in the Lagrange formalism.

To cope with data that is not linearly separable, the vectors **u** are mapped into a higher-dimensional space $\mathcal{F}$ by a map function Φ. In the space $\mathcal{F}$, the linear SVC algorithm is applied. The linear classifier in the space $\mathcal{F}$ will create a non-linear decision boundary in the original input space (Figure 2.3).

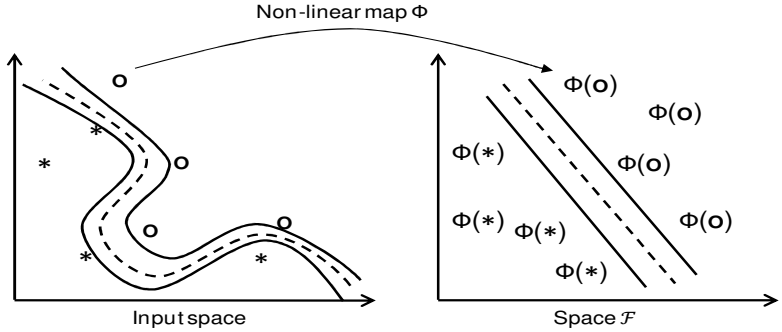

Figure 2: The SVC algorithm is applied in a high dimensional space (Modified redrawn from Cristianini and Schölkopf , 2002, p. 40)

The implementation of the SVC algorithm in the space $\mathcal{F}$ is done by using kernels; this consists of replacing the scalar product between training data with a kernel function in the formulation of the SVC algorithm. The kernel is a function in the input space of the vectors **u**, which returns the dot products of the images in some space $\mathcal{F}$, without even knowing the form of the map Φ:

$$k(\mathbf{u}_\alpha, \mathbf{u}_\beta) = \left\langle \Phi(\mathbf{u}_\alpha), \Phi(\mathbf{u}_\beta) \right\rangle \tag{9}$$

SVC testing

It implies to use the decision boundary (7) to allocate a single category *s* to the unsampled location **u** according to the rule:

$$s(\mathbf{u}) = \begin{cases} s_1 & if \quad \left(\sum_{\alpha=1}^{n} i(\mathbf{u}_\alpha)\eta_\alpha K(\mathbf{u},\mathbf{u}_\alpha) + b \right) > 0 \\ s_2 & if \quad \left(\sum_{\alpha=1}^{n} i(\mathbf{u}_\alpha)\eta_\alpha K(\mathbf{u},\mathbf{u}_\alpha) + b \right) < 0 \end{cases} \tag{10}$$

where $K(\mathbf{u}, \mathbf{u}_\alpha)$ is a symmetric positive-definite matrix that contains the values of the kernel function.

The unsampled locations taken together are called the testing set. The percentage of unsampled locations misclassified by the decision boundary is called testing error or generalization error. The complement of the generalization error to add to unity is called *generalization accuracy*.

Implementation of SVM

In (Hsu, Chang, & Lin, 2009) it is proposed a work flow for basic implementation of SVM, it is as follows:

- Scale the data. Sometimes this step is considered as a weakness of the machine learning techniques. That is, the fact they cannot handle datasets in their original units. On the contrary, this leads to one of the major strengths of these kinds of techniques. SVM is able integrate many attributes in the process of modelling. In geostatistics the variable of interest is represented as a regionalized variable as a function of a position vectors $\mathbf{u}$, in the case of SVM it can be expanded to many additional attributes. The position vector $\mathbf{u}$ in SVM can also integrate additional characteristics of the variable of interest at such locations. For example, at a location $\mathbf{u}_i = [x_i, y_i, z_i]$ the categorical variable of interest represents a certain rock type, say 0 or 1. It is possible to add mineralogy information, weathering characteristics, etc., so that, the location vector $\mathbf{u}$ becomes high dimensional. One advantage of this is the resulting modelled variable is not only a function of the geometric position in the domain, but also of other geologic characteristics. Scaling the data means to put all the elements of the position vector $\mathbf{u}$ in a unique unit system. For implementation scales of -1 to 1, 0 to 1 are used.
- Consider RBG kernel $K(\mathbf{x}, \mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|^2}$. This kernel type is preferred because it only requires a γ parameter. During the selection of parameters the user has to see only a two dimensional map. Recall the other parameter used in SVM is the penalty C. In the case studies presented these parameter maps are analyzed in order to find the smaller generalization error for this case.
- Use cross-validation to find the best parameters of the RBG kernel. This prevents over fitting (Hsu, Chang, & Lin, 2009). This type of cross-validation is different from the traditional practice. It is usually referred to as k-fold cross-validation. It divides the training set is small groups of data of similar size, each set is tested using γ and C parameters with the remaining information, then the resulting accuracy is the percentage of successful classification. This is done for each combination γ and C parameters in order to fill SVM parameters map.
- With the selected parameters in the previous perform training. From the γ and C parameters map one pair is selected. This is the decision of the user. Whether it is required to honour all the dataset or get the smaller generalization error the selected pair of parameters is used to classify the rest of the domain. One of the advantage in terms of computational speed is that once the γ and C parameters were chosen getting the categories of the unsampled locations is very fast.

Compared to geostatistical techniques the most important differences are:

- SVM classification cannot reproduce easily small variability as in the case of SIS when a nugget effect component is part of the categorical variable variogram.
- SVM not necessarily honour the available dataset. Basically it, propose models and by the use of the generalization error says how similar to reality are such models.
- It is computationally efficient when compared to SIS. Once the γ and C parameters were chosen, for categorization it is not required inverting matrices or so. SVM samples in the high dimensional in which side of the separation hyperplane the unsampled locations fall.
- Implementation of SVM does not require modelling variograms as is the case of SIS. The results are data driven. It is only necessary to scale the variable attributes and that is all.

Case Studies

Two case studies are presented. The first one is a easy two dimensional case were reality is known. It is presented for showing the impact in the selection of the γ and C parameters. The second case is a real case of a porphyry copper mine in Chile. SVM is used to build a three dimensional model of one rock type.

A binary figure that represents a river is drawn. From it 120 samples are drawn randomly in order to reproduce the initial figure (see Figure 3). The location vector consists of only x and y coordinates.

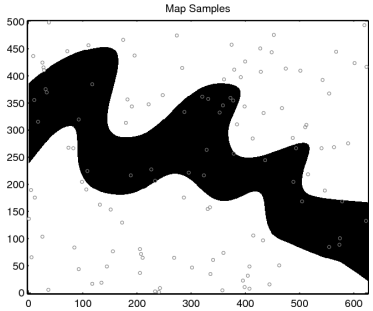

Figure 3: True figure map with 120 randomly sampled locations

A RBF kernel is considered with a range of the penalty parameter $log_2 C$ from -4 to 4 and $log_2 \gamma$ from 2 to 12. The selection of this range is recommended by (Hsu, Chang, & Lin, 2009). This is a good starting point to scan the response of the SVM with the presented data (see Figure 4). It can be seen that none of the SVM models can reproduce reality at 100%, the maximum reproduction reached is around 95%. This map (see Figure 4 bottom) can only be calculated when reality is known. In practice it is available the observed reproduction map (see Figure 4 top-left) and the *k*-cross validation map (see Figure 4 top-right). In the document the selection of the parameters are reduced to the combination of the two maps. That is, the target is to get the maximum cross validation accuracy and at the same time try to reproduce the major number of initial samples as possible. Other criteria in the selection of the parameters can be also implemented. For the different selection of the γ and C parameters different responses of the model is obtained (see Figure 5). Notice the map in right gets a better accuracy to reality despite one location is not honoured.

In a second study case one rock of the geology of a copper porphyry deposit in Chile in 3D is modelled (see Figure). The samples that correspond to the rock to be modelled are marked as black dots and of the different rock type in gray empty dots. The range of $log_2 \gamma$ and $log_2 C$ parameters are from -5 to 8 and 0 to 13 respectively. Notice that, so far the process consists of scaling the input dataset and selecting the range of the parameter maps. It is out of the focus of this paper to discuss about the criteria for an optimal selection of the γ and C parameters, therefore a reasonable pair is chosen. It is, $log_2 \gamma = 5$ and $log_2 C = 8$. Once the parameters are chosen it does not require too much time to generate the 3D rock type model (see Figure). Four slices of the solid are plotted in order to show the consistency with available dataset. Notice in some of the slices the available dataset is not honoured, however, the modelled rock type is consistent with the rest of the data.

Bibliography

Boser, B. E., Guyon, I., & Vapnik, V. (1992). A Training Algorithm for Optimal Margin Classifier. *Fifth Annual Workshop on Computational Learning Theory* (pp. 144-152). ACM Press.

Cortes, C. (1995). *Prediction of Generalization Ability in Learning Machines - PhD Thesis.* Rochester NY: Department of Computing Science, University of Rochester.

Cortes, C., & Vapnik, V. (1995). Support Vector Networks. *Machine Learning* , 273-297.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2009, May 19). *A Practical Guide to Support Vector Classification.* Retrieved July 1, 2009, from National Taiwan University, Department of Computer Science and Information Engineering: www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Muller, K. R., Smola, A., Ratsch, G., Scholkopf, B., Kohlmorgen, J., & Vapnik, V. (1997). Predicting Time Series with Support Vector Machines. *Artificial Neural Networks ICANN'97* (pp. 999-1004). Berlin: Springer.

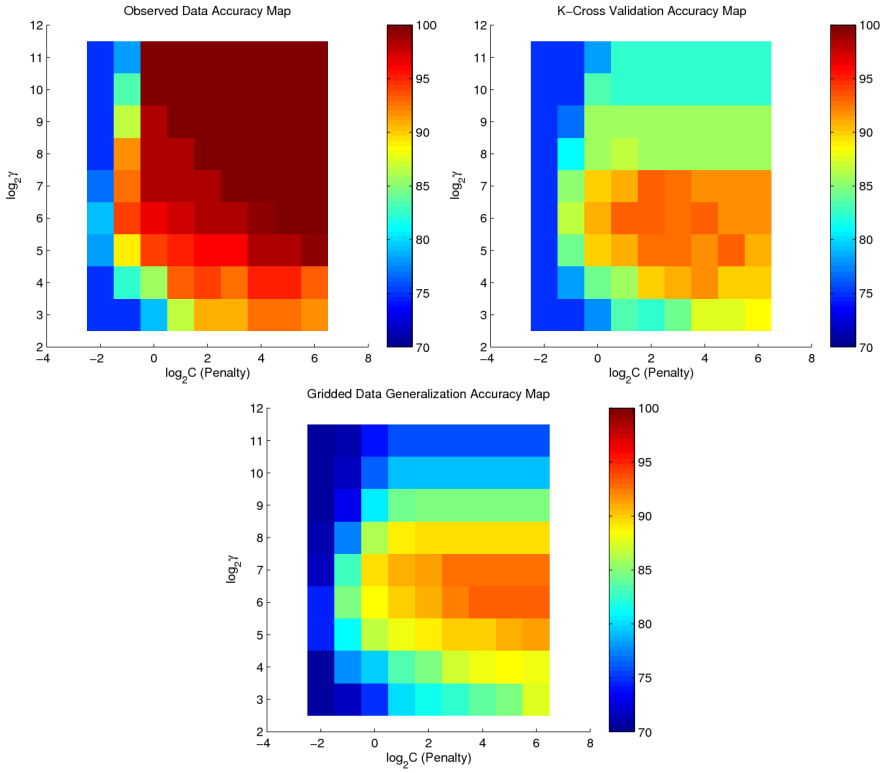

Figure 4: SVM parameters maps of accuracy compared to the input data (top left), k-fold cross validation (top right) and of reality (bottom)

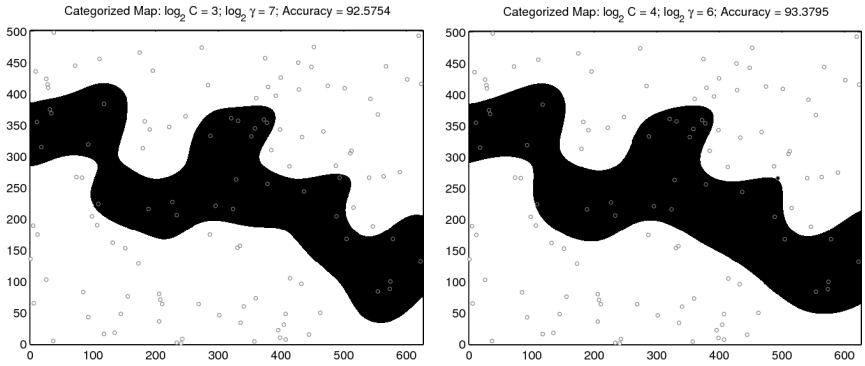

Figure 5: Two SVC maps with different penalty and gamma parameters

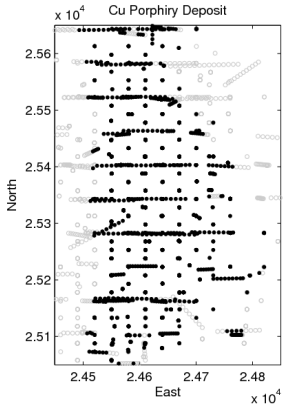

Figure 6: Exploratory Information of a Porphyry Deposit

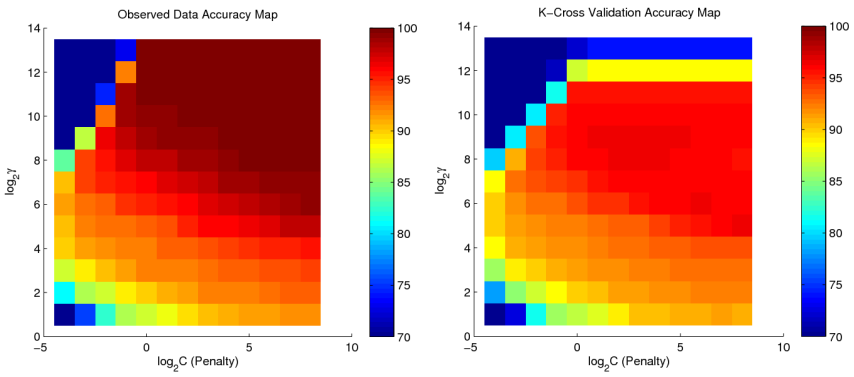

Figure 7: SVM map of accuracy compared to observed data (left) and k-fold cross validation (right)

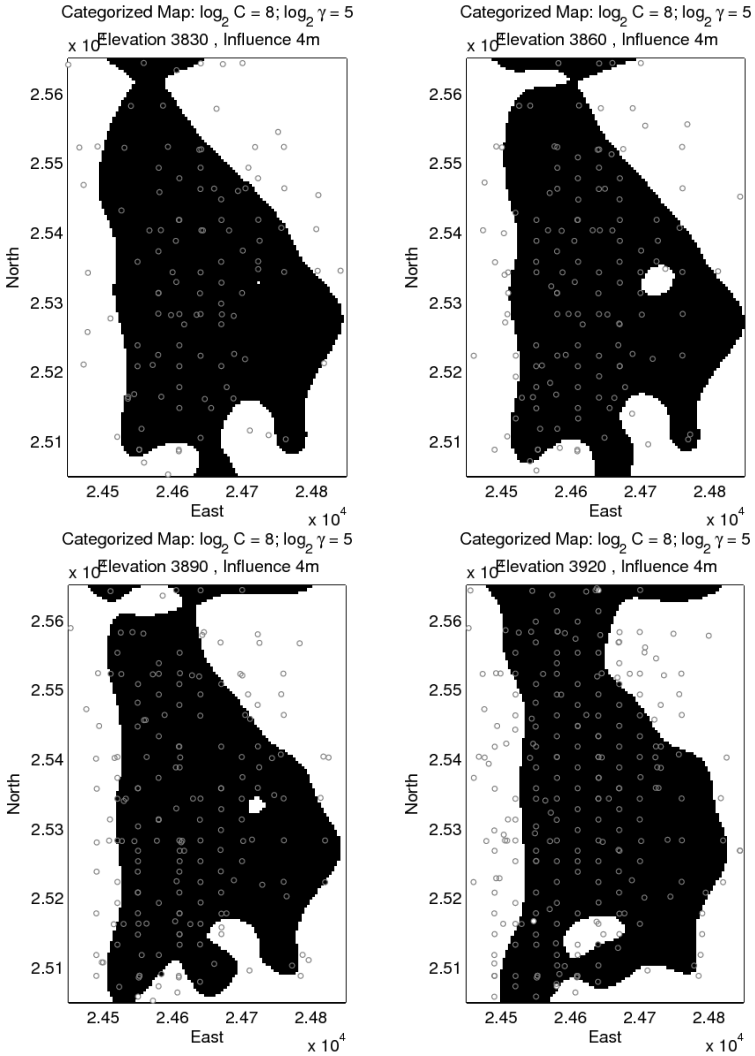

Figure 8: Four slices of a three dimensional volume calculated using SVM classification