Heuristic Support Vector Classification for Categorical Data

Enrique Gallardo and Oy Leuangthong

The Support Vector Classification (SVC) algorithm is a learning machine intended to solve the classification problem, which in the context of spatially-distributed data is the problem of assigning a single category (e.g. facies or rock types) to unsampled locations based on a limited set of observed data. The SVC response is determined by a pair of parameters (P, γ) that must be simultaneously selected by the user. The task of selecting the "best" pair (P, γ) is often made using cross-validation. This report proposes a novel and easy-to-implement technique to select a pair of SVC parameters that produce a response with good generalization ability, correct classification of all the observed data, and good reproduction of the global proportions.

Introduction to SVC

In the context of spatially-distributed data, the Support Vector Classification (SVC [Boser et al., 1992]) algorithm can be used to solve the problem of assigning a single category (e.g. rock types) to an unsampled location based on a limited set of observed data. For instance, consider Figure 1 where a single rock type, white or black, must be allocated to the unsampled location **u** based on the information collected from 10 locations.

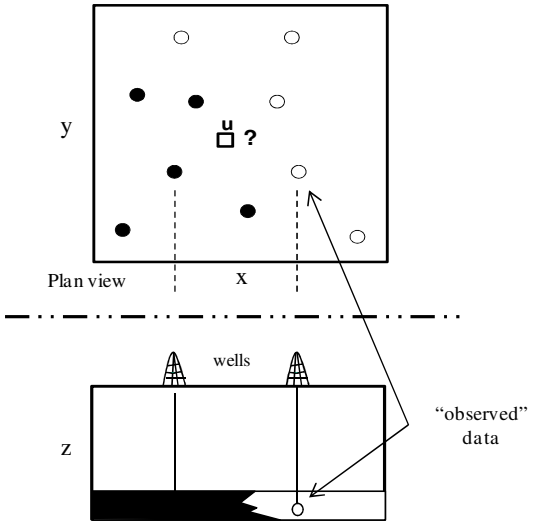

Figure 1: Classification problem. A single rock type, white or black, must be assigned to the unsampled location **u** based on 10 observed data points.

The SVC algorithm solves this problem in two steps: first, a boundary that separates the white and the black locations is found *only based on the sampled data*; then, this boundary is used as a classifier to assign a single rock type to the unsampled location **u** (Figure 2).

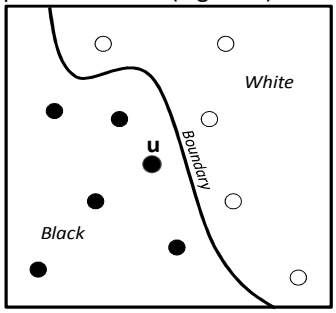

Figure 2: A boundary classifier assigns a single rock type to the unsampled location **u**.

The goal of the SVC algorithm is to find the "optimal" boundary. Here, optimality is measured by the *generalization accuracy*, which is defined as the percentage of locations (sampled and unsampled) in the domain of interest correctly classified by the SVC boundary.

$$Generalization\ accuracy = \frac{Number\ of\ \ sampled\ and\ unsampled\ locations\ correctly\ classified}{Total\ number\ of\ \ locations\ in\ the\ domain\ of\ \ interest} \cdot 100$$

In practice, the *generalization accuracy* cannot be established; we can only calculate the percentage of sampled locations correctly classified by the SVC boundary, which is called the *empirical accuracy*.

$$Empirical\ accuracy = \frac{Number\ of\ \ sampled\ locations\ correctly\ classified}{Total\ number\ of\ \ sampled\ locations} \cdot 100$$

The SVC algorithm states that the boundary separating the data with the maximum margin will also exhibit the highest generalization accuracy. Figure 3 illustrates this concept; intuitively, one can expect that the boundary line with a large margin in the right graph will have a better performance classifying unsampled locations than the boundary with a small margin in the left graph (Kecman, 2001, p.150).

Note: It might be necessary to misclassify some observed data in order to obtain a boundary with a large margin. In Figure 3, the boundary with the larger margin misclassifies one location while the boundary with the smaller margin correctly classifies all the given locations.

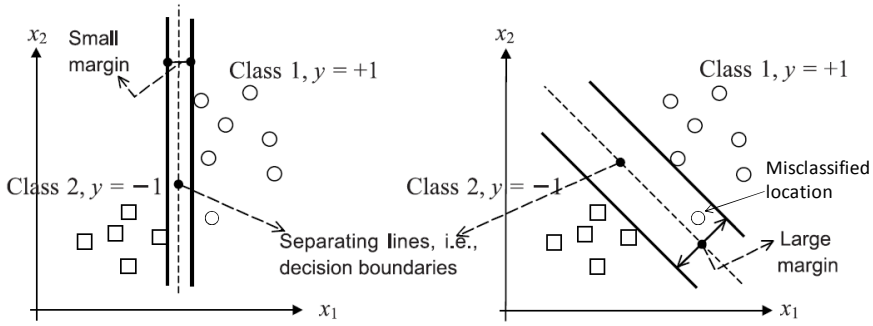

Figure 3: Two classifiers: a good one with a large margin (right) and a less acceptable one with a small margin (left). (Modified from Kecman, 2001, p.149)

The SVC boundary by construction has the form $\mathbf{w}^T\mathbf{u} + b = 0$, where **u** represents the locations and **w** and b are the parameters to be found when solving an optimization problem that balances the size of the margin and the number of misclassified locations. As described in detail in Gallardo and Leuangthong (2009a), solving this optimization problem requires the user to simultaneous select two parameters, *P* and γ. The former is a penalty parameter that controls the trade-off between margin and misclassifications, while the latter is the kernel parameter that produces a non-linear classifier. This paper proposes a novel heuristic technique to select the pair of parameters (P, γ).

Model selection

Traditional techniques: k-fold and leave-one-out (LOO) cross-validation

Applying the SVC algorithm with a Gaussian radial basis function (Grbf) kernel implies the simultaneous selection of the pair of parameters (*P*, γ), so that the boundary classifier can predict unsampled locations with the maximum generalization accuracy. Given that the true generalization accuracy is unknown, a proxy value is calculated by cross-validation.

A widely-used type of cross-validation is *k*-fold cross-validation. The observed data is randomly divided into *k* equal-sized subsets. Then, the SVC algorithm is sequentially trained using the *k*-1 subsets and tested in the remaining subset. Training is repeated *k* times and the percentage of data correctly classified for all the *k* subsets that are not included in the training data is recorded as the cross-validation accuracy (Abe, 2005, p. 73). When *k* is equal to the number of observed data, the technique is called leave-one-out (LOO) cross-validation.

To select the optimal pair (*P,* γ), the conventional approach calculates the *k*-fold or LOO cross-validation accuracy for every pair (*P,* γ) on a predefined grid-search and it chooses the one with the maximum value. To explore a wide range of parameter combinations, the grid is designed as an exponentially-growing sequence of *P* and *γ* values (Hsu, Chang and Lin, 2008). For instance, $P = \{2^{-3}, 2^{-2},\ldots, 2^{7}, 2^{8}\}$ and $\gamma = \{2^{-10}, 2^{-9},\ldots, 2^{10}, 2^{11}\}$.

The selected pair of parameters (*P,* γ) is used to train the SVC algorithm with the complete set of observed data. A theoretical description of cross-validation along with a benchmark of other traditional techniques for model selection can be found in Anguita, Boni, Ridella, Rivieccio and Sterpi (2005).

Heuristic selection of SVC parameters (P, γ)

Instead of cross-validation, this essay proposes to select the SVC parameters based on desirable geostatistical properties, that is, correct classification of the sampled data set and good reproduction of the global proportions of categories. The SVC solution obtained by this method is expected to exhibit good prediction property. This heuristic parameter selection process, as outlined in Figure 4, has the following steps:

1. Select a grid-search for the pair of parameters (P, γ)
2. Visit the nodes of the (P, γ) grid-search and at each node:
    a. Train the SVC algorithm using the observed data and calculate the empirical accuracy.
    b. Test the SVC algorithm using all the locations in the domain of interest and the model obtained in step (2a). Calculate the proportions on the resulting categorical map.
3. Plot contour lines of the empirical accuracy and the proportions over the (P, γ) grid-search. Select the node (P, γ) where the first contour of empirical accuracy of 100% intersects the closest contour to the target proportions.

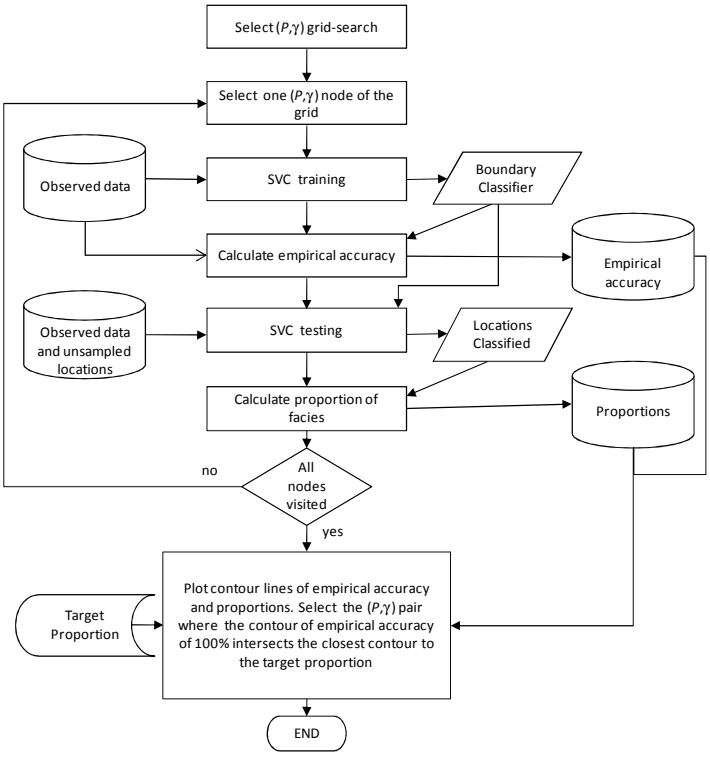

Figure 4: Work flow for the proposed technique to select the SVC parameters (*P,* γ)

The following case study illustrates the application of the new technique to solve a classification problem and compare the results to those obtained by the conventional approaches, ordinary indicator kriging and SVC with cross-validation.

Case study

A synthetic classification study illustrates the application of the proposed methodology. A reference data set of two rock types was generated and a subset was sampled to be used as observed data. Initially, the conventional ordinary indicator kriging (OIK) and SVC with k-fold cross-validation are used to allocate the rock types at the unsampled locations of the reference data; then, the proposed technique is applied. The results are compared and discussed.

Reference data and samples: The reference data set (Figure 5 left) is generated by unconditional Gaussian simulation using an isotropic spherical semivariogram model without nugget effect and a range of 200 m. The Gaussian field is transformed to a categorical variable of two rock types - white (code 1) and black (code 0) - using a threshold of 0 normal units. The noise or small scale variability in the map is removed by twice applying a moving window-cleaning procedure.

The reference data has a resolution of 10 m x 10 m spacing that spans an area of 1km x 1km, so it contains 10000 nodes. The proportions for the white and black rock types are 45.75% and 54.25%, respectively. A relatively large (2.25% of the reference data) sample was drawn to reduce the subjectivity in the construction of the geostatistical model, specifically to facilitate the modeling of the semivariogram and the calculation of the representative global proportions. The observed data is sampled from the reference map - nominally at 70 m x 70 m spacing. Figure 5 (right) shows the locations of the 225 samples.

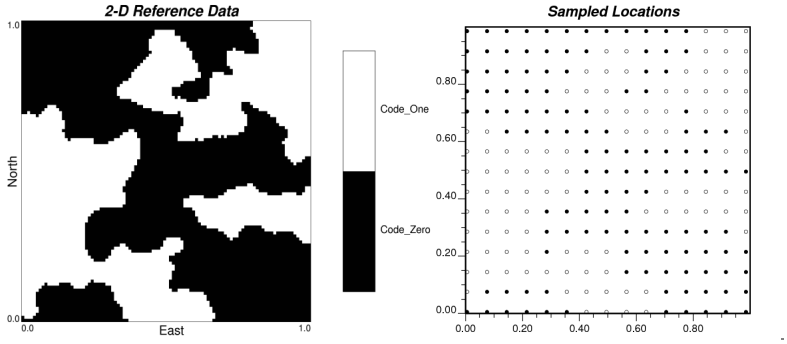

Figure 5. Reference map (left) and 225 samples to be used as "observed data" (right)

Classification by indicator kriging: The conventional ordinary indicator kriging was the first method applied to solve this problem. The exploratory data analysis indicates that the sample percentages for the white and the black rock types are 45.89% and 54.11%, respectively. These values are representative of the population. The standardized experimental indicator semivariogram was modeled as:

$$\gamma(\mathbf{h}) = Sph_{\substack{h\max=0.28 \\ h\min=0.19}}(\mathbf{h}).$$

Figure 6 (left) shows the map of conditional probabilities of occurrence for the white rock type at the resolution of the reference data set obtained by ordinary indicator kriging (OIK). For the binary problem at hand, a threshold rule that allocates the locations to the rock type whose probability of occurrence is greater than 50% was applied. The resulting categorical map (Figure 6 right) is compared node-to-node to the reference map to calculate the generalization accuracy which was 91.51%. The estimated map has proportions for the white and black rock types of 44.86% and 55.14%, respectively.

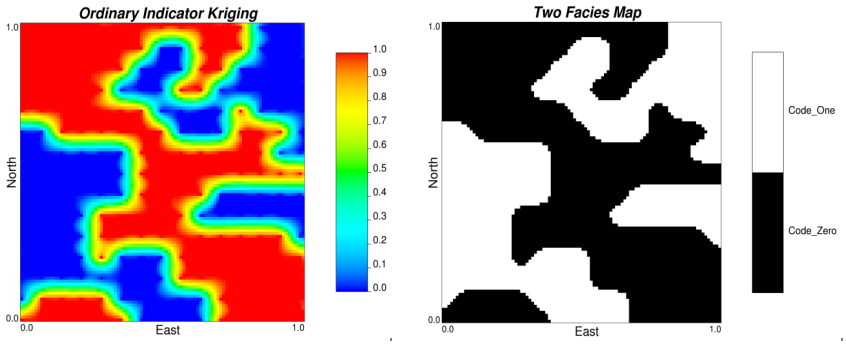

Figure 6: Ordinary indicator kriging map for the white rock type (left) and the resulting map of rock types after applying a threshold rule of 50% (right)

SVC with cross-validation: The conventional SVC requires perform model selection (to select the kernel and the parameters (P, γ)) and train and test the algorithm.

Usually, before model selection, the facies are coded as -1 and 1; to avoid numerical difficulties, the coordinates are linearly rescaled to the range [0, 1]. In this case, the reference data is already in the recommended scale. The Grbf kernel is chosen for this example and the pair of parameters (P, γ) is tuned as follow:

(1) An 81x81 grid-search space for the pair of parameters (P, γ) was defined to do *k*-fold cross-validation and *k* was set to 10. The sequence of values of the grid is: $P = \{2^{-2}, 2^{-1.9}, \ldots, 2^{-5.9}, 2^{-6}\}$ and $\gamma = \{2^{3}, 2^{-3.1}, \ldots, 2^{10.9}, 2^{11}\}$,

(2) For each pair (P, γ) in the grid-search the *k*-fold cross-validation accuracy is calculated. The pair with the highest accuracy is selected to train the SVC algorithm. Figure 7 shows the contour lines of the cross-validation accuracy, where the maximum value is 91.56% at the pair ($\log_2 P$, $\log_2 \gamma$) = (-0.1, 7.7).

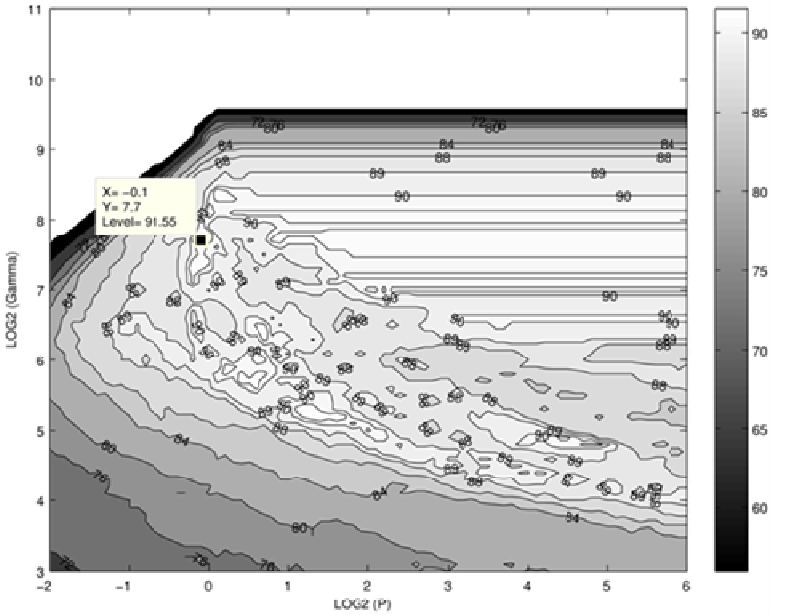

Figure 7: Contour lines of the cross-validation accuracy. The flag shows the maximum value of 91.56% at the pair ($\log_2 P$, $\log_2 \gamma$) = (-0.1, 7.7).

Once the pair (-0.1, 7.7) is selected, the SVC algorithm is trained using the 225 sampled locations. The result is a boundary classifier that is used to classify all the locations (sampled and unsampled) in the domain of interest. Figure 8 shows the categorical map obtained after the testing procedure. The map has a generalization accuracy of 91.37%. The percentages of white and black rock types are 44.5% and 55.5%, respectively.

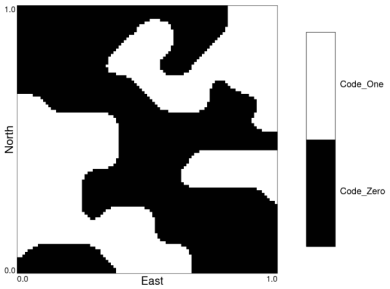

Figure 8: Map of rock types obtained by SVC with k-fold cross-validation

SVC with heuristic parameter selection. The proposed heuristic technique differs from the conventional SVC in the criteria used to select the parameters (*P*, γ). Using the same 81x81 grid-search previously defined, the key aspects of the procedure proposed in Section 2.2 are sketched in Figure 9. First, the empirical accuracy contour map (Figure 9a) and proportion contour map (Figure 9b) are plotted. Second, the maps are combined to find the intersection point between the 100% empirical accuracy contour line and the target proportion of 45% contour line (Figure 9c). In this case, the intersection was found at the pair (1.6, 6.9) which is used to train the SVC algorithm. Figure 10 shows the real combined contour map generated for this study case. The node flagged at ($\log_2 P$, $\log_2 \gamma$) = (1.6, 6.9) is the intersection between the empirical accuracy contour of 100% and the contour of 45.15% white proportion, which is the nearest value to the target 45.89% proportion.

Once the pair (1.6, 6.9) is selected, the SVC algorithm is trained using the 225 sampled locations. The resulting boundary classifier is used to generate the map of rock types shown in Figure 11. This estimated map has a generalization accuracy of 91.46%. The proportions of the white and black rock types are 45.15% and 54.85%, respectively.

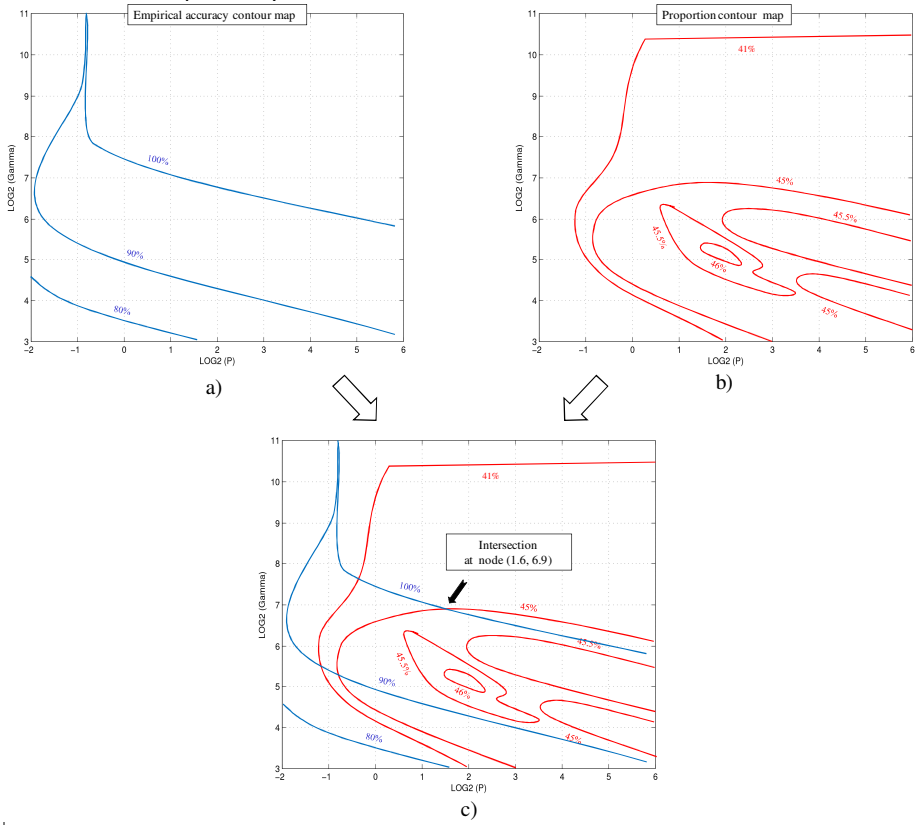

Figure 9: Empirical accuracy contour map (a), proportion contour map (b), and combined map (c) to find the intersection point between the 100% empirical accuracy contour line and the target proportion of 45% contour line.

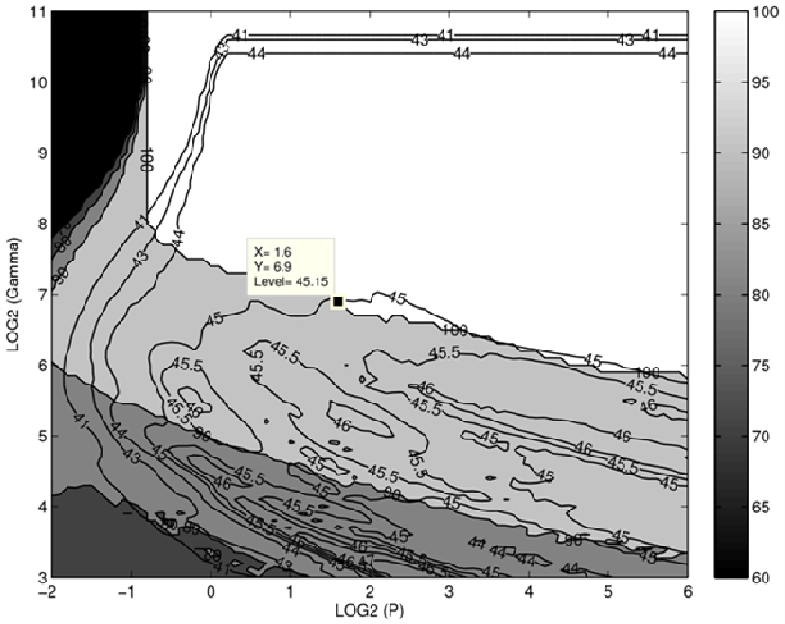

Figure 10: Contour lines of the empirical accuracy (grey scale) and the proportions of white rock type of the estimated categorical maps. The flag shows the intersection node ($\log_2 P$, $\log_2 \gamma$) = (1.6, 6.9).

Figure 11: Map of rock types obtained by the proposed technique.

Summary of numerical results and discussion. Figures 6, 8 and 11 and Table 1 show that the three methodologies produce similar numerical results, which is a direct consequence of having a relatively large amount of regularly-spaced data.

Comparing the conventional SVC to the proposed heuristic SVC approach, the latter not only has slightly better generalization accuracy, but fewer support vectors which is a desirable sparse property for the SVC classifiers.

The conventional geostatistics OIK technique produces slightly better generalization accuracy than the conventional and the proposed SVC algorithms. However, the SVC algorithm has the advantage of being totally automatic. As they do not require any subjective modeling decisions, the results are fully reproducible based on only the observed data.

The example presented here confirms that the responses of the OIK algorithm and the proposed SVC algorithm tend to converge, at least, for large data sets.

Table 1: Summary of numerical results of case study

Technique	Generalization accuracy (%)	Proportions of rock types (%)		Number of support vectors
		White	Black	
Reference data	N.A	45.75	54.25	N.A
Observed data (225 samples)	N.A	45.89	54.11	N.A
Geostatistical OIK	91.51	44.86	55.14	N.A
Heuristic SVC	91.46	45.15	54.85	137
k-fold SVC	91.37	44.5	55.5	216

N.A: Not apply

Conclusions

This paper introduced a methodology that is anchored in geostatistical criteria to select the SVC pair of parameters (*P,* γ). The case study showed that the proposed technique offers a solution to the classification problem that is comparable with those obtained by conventional ordinary indicator kriging. The new method has the advantage of being fully automatic and it is very easy to implement for modeling of rock types. The synthetic example also showed that the proposed technique over-performed the traditional SVC with *k*-fold cross-validation in terms of generalizations accuracy and reproduction of global proportions.

References

1. Abe, S. (2005) Support vector machine for patter classification. Springer, USA.
2. Boser, B.E., Guyon, I., and Vapnik, V. (1992) A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144-152. ACM Press.
3. Burges, C.J. (1998) A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery 2, pp. 121-167. Kluwer Academic Publishers, Boston.
4. Chang, C.-C. and C.-J.Lin. (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. Retrieved 22 December 2008.
5. Deutsch, C.V., and Journel, A.G. (1998) GSLIB: geostatistical software library and user's guide. Oxford University Press, New York.
6. Hsu C-W., Chang Ch-Ch., and Lin C-J. (2008) A practical guide to support vector classification. Department of Computer Science. National Taiwan University, Taipei 106, Taiwan. http://www.csie.ntu.edu.tw/~cjlin. Last updated: May 21, 2008.
7. Kecman, V. (2001) Learning and softcomputing. The MIT Press.
8. Vapnik, V. (1995) The nature of statistical learning theory. Springer-Verlag, New York.
9. Gallardo, E., and Leuangthong, O., (2009a) An Introduction to Support Vector Classification for Geostatistical Applications. CCG Report 2009.