Classification for the Determination of Estimation Domains in a Cu-Zn Skarn Deposit in Central Peru, New Approach using Gaussian Kernel Support Vector Machine

Harold Velasquez, Marjory Aguilar.

Abstract— The Ore-control block-model in an open pit mine constitutes the final outcome after rigorous analysis and interpretations of a wide range of geological data. Moreover, modeling different variables in current tools such as GIS software or specialized programs may be highly time consuming and these software's and tools may restrict you to use determined types of data and can be un-accurate when they are utilized for making attempts to find out multivariate relationships. One of these ore-control tasks that has to be done is the determination of the short-term grades for the block-model, within which diverse mathematical calculations are carried out. In order to get the grade estimation, geologists require to determine the Estimation Domain for every lithology and then go forward with the estimation techniques using the laboratory grades from blastholes samples as an input, so it is clear that estimation domains are essential for ore-control purposes. Estimation domains require logged lithology and grades input of every blast hole sample; the logged lithology is directly obtained by the geology staff by describing the detritus from blastholes. This paper aims to present novel results in determining the relationships of multivariate laboratory assays in a Cu-Zn Skarn deposit and its corresponding logged lithology using kernel support vector machine algorithm, so with this, it may be possible for geologists to forecast lithology of every sample, based on its chemical content, and so, they will be able to determine estimation domains.

Index Terms— Blasthole, Kernel trick, LaGrange multiplier, Litho-geochemistry, Radial Basis Function, Skarn, Support Vector Machine.

1 Introduction

Support Vector Machines (SVM) recently became one of the most popular classification methods. They have been used in a wide variety of applications such as text classification [1], facial expression recognition [2], gene analysis [3] and many others. In many real-life situations we want to be able to assign an object to one of several categories based on some of its characteristics. For example, based on the results of several laboratory assays we want to be able to say whether a blast hole detritus sample belongs to a particular estimation domain or not [4]. In this paper we address Vapnik's [5] Support Vector Machine (SVM), and in particular its training when the so-called Gaussian kernel is adopted [6].

Support Vector Machines can be thought of as a method for constructing a special kind of rule, called a linear classifier, in a way that produces classifiers with theoretical guarantees of good predictive performance (the quality of classification on unseen data). The theoretical foundation of this method is given by statistical learning theory [7]. Literature study reveals that SVM has been used to solve classification problems in different domains of research such as hyperspectral imaging [8], reservoir characterization [9], imbalanced dataset [10], and machines [11]. The theory and approaches of SVM (binary or multiclass) has been discussed in length [12]. In essence, SVM is a binary classifier. Therefore, in case of a multiclass problem as in our case of study, the problem is divided into a series of binary problems which are solved by binary classifiers, and finally the classification results are combined following either the one-against-one or one-against-all strategies [13].

To start comprehending this work it is required we understand the detritus logging. Detritus logging is a significant labor carried out by the geology staff on a daily basis. This, consists on describing the mineralogical features, the approximate visually determined ore content and besides, some physical properties of the detritus coming from the blast holes. Each and every day over 300 blastholes samples are logged and the most important variable described is the lithology which is afterwards utilized for the grade estimation process. As the time passes, activities need to be optimized according to the trending and so automatic data generating need to impose over old methods. Litho-geochemistry; which is defined as the determination of the chemical composition of the rock units with the objective of detecting distribution patterns of elements that are spatially related, might help us; in this case, a skarn zonation. In other words, litho-geochemistry can be used for the classification of particular types of rocks and zones based on chemical elements content.

In this work we are seeking the results of applying classification with Support Vectors Machine on a specific set of data available from a Skarn deposit in Central Peru. This is not a determinant result due to the set of data do not contain the specific elements analyzed that would be needed in order to characterize a specific type of rock, however it will give us initial impressions of how effective the SVM method is, and if the current set of elements assayed in this deposit are sufficient to determine estimation domains .

2 LITERATURE REVIEW

2.1 Support Vector Machine

SVM's are one of the top edge supervised learning algorithms. To gain a better grasp of how SVM works we need to first get an insight of what margins is, and the idea of separating data with a widest space between them. We also need to understand the Lagrange duality that we are not going to discuss in depth here. Finally, kernels will show us a way to apply SVM's in a higher dimensional space. Many kernels have been developed for special applications such as sequence matching in bioinformatics [14] and that is the reason because general properties of kernels are described in many publications, including [15]. The main objective of SVM is to find the optimal hyperplane in order to linearly or no- linearly separates the data points in two or more components by maximizing the so-called margin.

2.2 SVM Formulation

It is better to first understand the support vector machine concept with binary classes due to it can be later extrapolated for multi-classes analysis.

TABLE 1

Variables used in Schematization of SVM Method

Symbol	Definition
	vector normal to the hyperplane
Ȁ| ||	perpendicular distance from hyperplane to origin
ؠ	input variable
ଢā ଗ	distances from H1 and H2 to hyperplane, respectively

In the image below we can see the two classes of points and the hyperplane passing in the midst of these two-point classes in such a way to give us the widest road or gap to separate the two group of points displayed. This is known as the *widest street approach* and represents the best intuitive way to separate two groups of data.

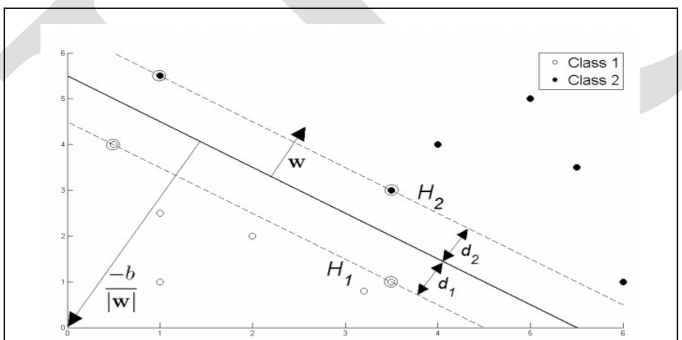

Fig. 1. Schematization of the SVM algorithm.

The formulation is as follows:

$$w\,.\,x_i+b\geq 1\ for\ y_i=+1 \qquad (1.1)$$
$$w\,.\,x_i+b\leq -1\ for\ y_i=-1 \qquad (1.2)$$

Combining above two equation, it can be written as:

$$y_i(w\,.\,x_i+b)-1\geq 0\ for\ y_i=+1,-1 \qquad (1.3)$$

Now the circled points represent the *Support Vectors* which lie closest to the hyperplanes H1 and H2, and so these can be described by:

$$x_i.\,w+b=+1\ for\ H_1 \qquad (1.4)$$
$$x_i.\,w+b=-1\ for\ H_2 \qquad (1.5)$$

As it can be seen, we get two planes corresponding to H1 and H2 which are the margins and M meaning the distance between these margins, so in order to maximize the distance between them it equals the operation of maximizing 1/ | w |

$$M=\left(\frac{1-b}{|w|}\right)-\frac{-1-b}{|w|}$$

(2.1)

$$M=2/|w| \qquad (2.2)$$

which is the same to minimize the term:

$$min\|w\|^2/2\ ;\ y_i(w\,.\,x_i+b)-1\geq 0\ for\ i=1\ldots l \qquad (2.3)$$

So, in order to solve the equation taking into account the constraints in this minimization, we need to allocate them Lagrange multipliers α, where $\alpha i \geq 0\ \forall i$:

$$L_P=\frac{1}{2}\|w\|^2-\alpha\left[y_i\left(x_i.\,w+b\right)-1\forall_i\right] \qquad (2.4)$$

$$¿\frac{1}{2}\|w\|^2-\sum_{i=1}^{L}\alpha_i\left[y_i\left(x_i.\,w+b\right)-1\right] \qquad (2.5)$$

$$¿\frac{1}{2}\|w\|^2-\sum_{i=1}^{L}\alpha_i y_i\left(x_i.\,w+b\right)+\sum_{i=1}^{L}\alpha_i \qquad (2.6)$$

We wish to find the **w** and b which minimizes, and the α which maximizes (whilst keeping $\alpha i \geq 0\ \forall i$). We can do this by differentiating L_P with respect to **w** and **b** and setting the derivatives to zero so we get the following final equations:

$$w=\sum_{i=1}^{L}\alpha_i\,.\,y_i\,.\,x_i \qquad (2.7)$$

$$0=\sum_{i=1}^{L}\alpha_i\,.\,y_i \qquad (2.8)$$

2.3 Gaussian Radial Basis Function (SVM)

In this section we will talk about the effects of varying the gamma and C parameters of the Radial Basis Function Kernel SVM from the Scikit library. And so we can say that a large C gives you low bias and high variance; low bias due to you penalize the cost of misclassification a lot; on the contrary, a small C gives you higher bias and lower variance. On the other hand, small gamma will give you low bias and high variance while a large gamma will give you higher bias and low variance. The task is related to find the best C and Gamma hyper-parameters which are usually found using Grid-Search analysis. Proper choice of C and gamma is critical to the SVM's performance.

This needs to be taken into account that C parameter does not appear directly within the RBF kernel equation, since it is the penalty associated to the instances which are misclassified or violates the maximal margin. It is also very important to know that when performing SVM classification, it's often helpful to scale the training data for SVM training, for example by subtracting the mean and dividing by the standard deviation, and afterwards scale the test data with the mean and standard deviation of training data. The next equation represents the Gaussian Radial Basis Function used in this paper.

$$K(x, x') = \exp(-\nu \|x - x'\|^2), \nu > 0 \qquad (3.1)$$

3 Data Description

3.1 Data

Samples were gathered from a same zone of the deposit studied, not all existing samples in that region were considered due to the high computer resources it would take to process them. As a result, 100 samples were taken for each of the 5 rock types or classes, summarizing a total of 500 data for the purpose of the present paper. Laboratory assays were given for the 500 blastholes samples in elements such as Copper, Zinc, Bismuth, Molybdenum, Arsenic, Lead, Silver, Iron, Cobalt and Sulfur. The 500 samples set of data also contains a field with a code of rock type which was determined by geology staff. In this study we have only considered five main lithologies corresponding to the endo-skarn and exoskarn in such a way that the calculations may give us representative results, these are the rocks corresponding to the intrusive; a non-Cu-rich proximal endo-skarn; a Cu-rich endo-skarn; and 2 rocks in the proximal exoskarn. We will now name them in a range from Class 1 to Class 5 respectively for later analysis.

3.2 Data Statistics

Finding the performance of Gaussian RBF Kernel SVM using the best fit parameters is what we are going to develop in the next paragraphs, we are not going to deepen into the Exploratory data analysis for our 500 samples data set since that is out of the scope of this research.

Additionally, in order to keep data secure we will only present a brief summary of the statistics of the used data. In the tables below we can see the statistics of the data used. The basic statistics are presented for the ten elements assayed.

TABLE 2
Statistics of the whole set of data

	CU_PCT	ZN_PCT	BI_PPM	MO_PCT	AS_PPM	PB_PCT	AG_PPM	FE_PCT	CO_PPM	S_PCT
count	500.000000	500.000000	500.00000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000
mean	0.982580	0.922740	11.39200	0.043194	29.880000	0.036780	10.698000	11.591000	27.118000	6.452720
std	0.829547	2.217569	32.14132	0.046334	31.116395	0.172989	10.073418	7.404849	21.716706	6.673124
min	0.020000	0.010000	3.00000	0.001000	10.000000	0.010000	1.000000	0.780000	5.000000	0.150000
25%	0.370000	0.020000	3.00000	0.005000	10.000000	0.010000	4.000000	4.882500	10.000000	1.227500
50%	0.840000	0.070000	3.00000	0.028000	20.500000	0.010000	8.000000	11.660000	22.000000	3.845000
75%	1.352500	0.682500	8.00000	0.073000	36.250000	0.020000	15.000000	17.057500	38.000000	10.132500
max	6.100000	17.400000	440.00000	0.314000	400.000000	3.160000	85.000000	30.580000	169.000000	35.810000

On the upcoming tables the summarized statistics is presented by each rock type or classes from 1 to 5, so we can get a better idea of the chemical content they have.

TABLE 3
Basic statistics of Class 1

	CU_PCT	ZN_PCT	BI_PPM	MO_PCT	AS_PPM	PB_PCT	AG_PPM	FE_PCT	CO_PPM	S_PCT
count	100.000	100.000	100.000	100.000	100.000	100.00	100.000	100.000	100.000	100.000
mean	0.176	0.014	3.880	0.090	11.120	0.01	1.800	1.936	7.110	0.572
std	0.269	0.010	2.447	0.036	3.867	0.00	2.005	1.586	4.497	0.684
min	0.020	0.010	3.000	0.031	10.000	0.01	1.000	0.780	5.000	0.150
25%	0.030	0.010	3.000	0.069	10.000	0.01	1.000	1.140	5.000	0.240
50%	0.065	0.010	3.000	0.088	10.000	0.01	1.000	1.305	5.000	0.310
75%	0.143	0.010	3.000	0.102	10.000	0.01	1.000	1.785	7.000	0.452
max	1.210	0.080	17.000	0.295	43.000	0.01	10.000	8.430	29.000	3.540

Intrusive Rock. Waste rock, presents no mineralization in either copper or zinc.

TABLE 4
Basic Statistics of Class 2

	CU_PCT	ZN_PCT	BI_PPM	MO_PCT	AS_PPM	PB_PCT	AG_PPM	FE_PCT	CO_PPM	S_PCT
count	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000
mean	0.756	0.101	5.890	0.060	23.800	0.018	5.810	7.852	18.000	2.892
std	0.520	0.188	7.402	0.052	42.842	0.019	3.237	3.908	12.475	2.946
min	0.100	0.010	3.000	0.002	10.000	0.010	1.000	2.040	5.000	0.400
25%	0.378	0.020	3.000	0.022	10.000	0.010	4.000	4.845	10.000	1.218
50%	0.660	0.030	3.000	0.049	14.000	0.010	5.000	7.305	13.000	1.605
75%	0.940	0.062	6.000	0.082	22.000	0.020	7.000	10.455	22.250	3.465
max	2.840	0.950	60.000	0.314	400.000	0.100	20.000	20.840	61.000	15.890

Near and slightly developed endo-skarn rock type. Mostly with no or low content of mineralization in copper.

TABLE 5
Basic Statistics of Class 3

	CU_PCT	ZN_PCT	BI_PPM	MO_PCT	AS_PPM	PB_PCT	AG_PPM	FE_PCT	CO_PPM	S_PCT
count	100.000	100.000	100.000	100.000	100.00	100.000	100.000	100.000	100.000	100.000
mean	1.606	0.294	5.990	0.045	21.21	0.018	12.620	13.137	31.860	5.641
std	0.870	0.724	4.945	0.037	14.25	0.015	8.713	5.131	20.526	4.094
min	0.200	0.010	3.000	0.002	10.00	0.010	3.000	3.860	6.000	0.560
25%	1.072	0.040	3.000	0.021	11.00	0.010	8.000	9.620	20.750	3.030
50%	1.495	0.060	3.000	0.038	17.00	0.010	10.000	12.205	26.000	3.895
75%	1.940	0.132	8.000	0.052	27.00	0.020	16.000	16.275	39.000	7.798
max	6.100	4.120	26.000	0.202	102.00	0.090	64.000	26.460	169.000	21.690

Well-developed dark garnets rock type. With high concentration of copper. Mineralized zone.

TABLE 6
Basic Statistics of Class 4

	CU_PCT	ZN_PCT	BI_PPM	MO_PCT	AS_PPM	PB_PCT	AG_PPM	FE_PCT	CO_PPM	S_PCT
count	100.000	100.000	100.00	100.000	100.000	100.00	100.000	100.000	100.000	100.000
mean	1.184	0.841	6.08	0.007	43.920	0.02	13.440	19.571	46.150	13.729
std	0.702	1.590	6.08	0.012	25.846	0.02	7.929	4.961	20.753	6.832
min	0.240	0.020	3.00	0.001	14.000	0.01	3.000	9.020	8.000	1.040
25%	0.740	0.090	3.00	0.002	27.750	0.01	8.000	15.848	30.750	8.888
50%	0.960	0.205	3.00	0.003	39.000	0.01	11.000	18.985	42.500	13.375
75%	1.465	0.678	7.25	0.007	53.000	0.02	16.000	22.207	57.750	18.352
max	3.740	8.920	35.00	0.087	178.000	0.15	42.000	30.580	97.000	27.440

Near exo-skarn with mineralization of copper and zinc. This rock type present garnets are brownish and greenish coloring.

TABLE 7
Basic statistics of Class 5

	CU_PCT	ZN_PCT	BI_PPM	MO_PCT	AS_PPM	PB_PCT	AG_PPM	FE_PCT	CO_PPM	S_PCT
count	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000
mean	1.191	3.363	35.120	0.013	49.350	0.118	19.820	15.459	32.470	9.430
std	0.831	3.710	66.111	0.023	33.286	0.376	12.597	4.184	21.057	6.406
min	0.080	0.080	3.000	0.001	10.000	0.010	3.000	2.010	5.000	0.660
25%	0.648	1.065	3.000	0.003	29.000	0.010	12.000	13.995	16.000	4.275
50%	0.950	1.810	7.000	0.006	38.000	0.015	18.000	15.820	28.000	8.140
75%	1.568	4.010	34.750	0.013	51.250	0.042	23.000	17.762	42.000	13.152
max	4.400	17.400	440.000	0.157	161.000	3.160	85.000	28.290	115.000	35.810

Near exoskarn presented next to the Class 4 Rock type. This rock type presents higher Zn content mainly and mineralization of copper as well. Garnets are light-green coloured.

4 Methodology

5 Results and Analysis

In the evaluation carried out the best parameters found were C: 100.0, gamma: 0.1 with a score of 0.77

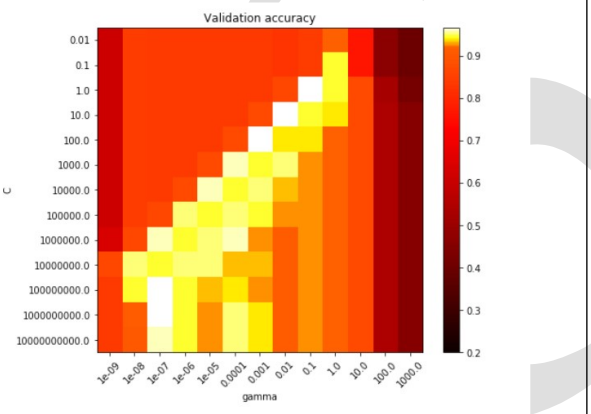

Fig. 2. Heat map of the classifier's cross validation accuracy as a function of C and gamma. In our example, the best parameters lie on the upper center of the grid.

After evaluating the algorithm, confusion matrix precision gives us the first sight of the effectiveness in the predictions, 0.20 percent of the whole data was used as the test size, it means 100 samples. In the figure below we can appreciate that the highest performances were achieved in Class 1 (intrusive) where only 1 sample was misclassified as Class 2 and 3 samples as Class 5; and in relationship with Class 5 there are no misclassifications. The weakest performance was obtained in the Class 4 where 22 samples were misclassified as Class 5 rock type.

```
[[21  1  0  0  3]
 [ 0 11  2  1  5]
 [ 0  1 10  0  7]
 [ 0  0  0  2 22]
 [ 0  0  0  0 14]]
```

Fig. 3. Confusion Matrix resulted from evaluation.

In the Classification report the best precision was reached in Class 1 rock type which turns to be the intrusive with no content of mineralization and the worst performance was encountered into the Class 4 Rock type with a precision of 0.67 and a recall value of 0.08, this class correspond to the near low developed endo-skarn.

TABLE 8
Final Classification Report

	Precision	Recall	f1-score	Support
1	1.00	0.84	0.91	25
2	0.85	0.58	0.69	19
3	0.83	0.56	0.67	18
4	0.67	0.08	0.15	24
5	0.27	1.00	0.43	14

The weighted average is 0.76 for precision and 0.58 for recall.

6 Conclusion

As a result of the evaluations, we conclude that the performance of RBF Kernel SVM on 500 samples from Skarn deposit was moderately acceptable so far aiming to help as a tool for the determination of the estimation domains, it seems the rock type classification based on chemical sign may represent a possible option if we account with a complete set of chemical element contents. It is recommended to perform similar calculations but regarding complementary data of the assays that were not taken into account in this work.

In any event, further calculations with different parameters range must be tested in the future with a higher amount of data set to enhance the results. With respect to the lithological logging, sometimes a blast hole sample presents more than a solely rock, in this study only the main rock was considered and that could negatively affect the results presented. Finally, it is necessary to evaluate the data with others Kernel SVM functions to contrast the results and to get a better comprehension of them.

References

[1] T. Joachims. "Text categorization with support vector machines: Learning with many relevant features". In Proceedings of the European Conference on Machine Learning. Springer, 1998.

[2] P. Michel and R. E. Kaliouby. Real time facial expresion recognition in video using support vector machines. In Proceedings of ICMI'03, pages 258–264, 2003.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3):389–422, 2002.

[4] Dmitriy Fradkin and Ilya Muchnik, "Support Vector Machine for Classification," DIMACS Series in Discrete Mathematics and Theoretical Computer Science: ResearchGate, pp. 227-236, 1989.

[5] V. Vapnik. The support vector method. In ICANN97, pages 263–271, 1997.
[6] B. Scholkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and 18 V. Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. IEEE Transaction on Signal Processing, 45:2758–2765, 1997
[7] V. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, 2nd edition, 1998.
[8] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," IEEE Trans. Geosci. Remote Sens., vol. 42, no. 8, pp. 1778–1790, 2004.
[9] A. Al-Anazi and I. D. Gates, "A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs," Eng. Geol., vol. 114, pp. 267–277, Aug. 2010.
[10] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," Mach. Learn. ECML 2004, vol. 3201, pp. 39–50, 2004.
[11] S. Fei, C. Liu, and Y. Miao, "Support vector machine with genetic algorithm for forecasting of key-gas ratios in oil-immersed transformer," Expert Syst. with Appl., vol. 36, pp. 6326–6331, 2009.
[12] A. Mathur and G. M. Foody, "Multiclass and binary SVM classification: Implications for training and classification users," IEEE Geosci. Remote Sens. Lett., vol. 5, no. 2, pp. 241–245, 2008.
[13] Soumi Chaki, Aurobinda Routray et al. "A novel multi-class SVM based framework to classify lithology from well logs: a real-world application". ResearchGate, 2016.
[14] T. Jaakkola, M. Diekhaus, and D. Haussler. Using the fisher kernel method to detect remote protein homologies. pages 149–158, 1999.
[15] N. Cristianini and J. Shawe-Taylor. An Introduction to support vector machines and other kernelbased learning methods. Cambridge University Press, 2000.