ISSN (Print) : 2319-8613
ISSN (Online) : 0975-4024 

# Vector support machines applied to the implementation of an automated bacterial classification system by taxonomy

Danilo A López Sarmiento [#1], Nelson E Vera Parra [#2], Oscar D Flórez Cediel [#3]

[# 1] Full Time Professor at Universidad Distrital Francisco José de Caldas,
Faculty of Engineering, Bogotá (Colombia-South America)
dalopezs@udistrital.edu.co

[#2] Full Time Professor at Universidad Distrital Francisco José de Caldas,
Faculty of Engineering, Bogotá (Colombia-South America)
neverap@udistrital.edu.co

[#3] Full Time Professor at Universidad Distrital Francisco José de Caldas,
Faculty of Engineering, Bogotá (Colombia-South America)
odflorez@udistrital.edu.co

***Abstract*— Among the challenges in the field of medical sciences are those related to the classification of bacteria by taxonomy; However, with the emergence of vector support machines, it is possible to optimize this task through automatization by separating classes in space called hyperplanes. In this sense, the article assesses performance when classifying bacteria using the quadratic, cubic and radial Gaussian kernel functions. The results obtained allowed to conclude preliminarily that the VSM implemented in the research did not feature the best performance for the data sequence entered into the system, reaching a maximum global performance in the classifier no greater than 30.25%; However, it is necessary to continue applying modifications to the model developed in order to determine the possibility of increasing the success percentage.**

***Keyword***— Multiclass classification, flowchart, kernel, VSM, taxonomy.

## I. INTRODUCTION

Vector support machines are a set of algorithms developed by Vladimir Vapnik that belong to the family of linear classifiers since they induce linear separators or hyperplanes in spaces with very high dimensional characteristics with a very particular inductive bias [1]. They can be used for classification, regression, density estimation, novelty detection, and other applications. In the simplest case, which is that of two classes, VSMs find a hyperplane that separates the two data classes with the greatest possible margin. This leads to good accuracy in the generalization of unnoticed data and supports specialized optimization methods that allow the VSM to learn from a large amount of data [2].

VSMs have a strong mathematical basis and are closely related to some well-established statistics theories. They not only try to classify train data correctly, but also maximize the margin to improve generalization performance. This formulation leads to a separation hyperplane that depends only on the (usually small part of) data points that are in the margin, which are called support vectors. Hence, the complete algorithm is called a vector support machine. In addition, since real-world data analysis problems often involve non-linear dependencies, VSMs can easily be extended to non-linearity models by means of semi-definite positive kernels. On the other hand, VSMs can be trained by quadratic programming, which facilitates the theoretical analysis, and makes designing efficient solvers that scale for large datasets more comfortable. Finally, when applied to real-world data, VSMs often deliver state-of-the-art performance in accuracy, flexibility, robustness, and efficiency [3]. Unlike Artificial Neural Networks (ANNs) which use the Empirical Risk Minimization (ERM) principle during the train phase, VSMs are based on Structural Risk Minimization (SRM), which has shown a better performance than ERM, since Support Vector Machines minimize an upper limit to the expected risk, unlike ERM that minimizes the train data error [4]. The VSM maps the entry points to a space of larger dimension characteristics, and then finds the hyperplane that separates them and maximizes the margin between the classes [5]. The mathematical formulation of Support Vector Machines varies depending on the nature of the data; that is, there is a formulation for linear cases and, on the other hand, a formulation for non-linear cases. It is important to clarify that, in a general classification way, support vector machines seek to find an optimal hyperplane that separates classes [5], [6].

In the last decades the classification (in the context of VSMs) has been extended to the multi-label case, structured output and multiclass [7]; and it is precisely in this last option (multiclass VSM) that the present article focuses, applying it to the categorization of taxonomic bacteria.

## II. MULTICLASS VSM

According to [7] For multiclass classification and structured output classification where the possible label established $Y$ can be large, maximum margin machines can be formulated through the introduction of a joint function map $\phi$ in pairs $(x_t, y)\ (y \in Y)$. Where $\Delta(y_t, y)$ is the discrepancy between the real label $y_t$ and the candidate label $y$, and the fundamental form can be written as [7]:

$$minimizar_{w,\xi_t} \frac{\lambda}{2}||w||^2 + \frac{1}{n}\sum_{t=1}^{n}\xi_t, \qquad (1)$$
$$s.t.\langle w, \phi(x_t, y_t) - \phi(x_t, y_t)\rangle \geq \Delta(y_t, y) - \xi_t, \forall t, y,$$

And the dual form is:

$$minimizar_{\alpha_{ty}} \frac{1}{2\lambda}\frac{\sum \alpha_{ty}\alpha_{t'y'}}{(t,y),(t',y')}\langle\phi(x_t, y_t) - \phi(x_t, y)\phi(x_{t'}, y_{t'}) - \phi(x_{t'}, y')\rangle - \sum_{t,y}\Delta(y_t y) \qquad (2)$$

$$s.t\ \alpha_{ty} \geq 0, \forall t, y; \sum_{y}\alpha_{ty} = \frac{1}{n}, \forall t. \qquad (3)$$

The kernelization can be carried out by simply replacing all internal products $\langle\phi(x_t, y)\phi(x_{t'}, y')\rangle$ with a joint kernel $k(\phi(x_t, y)\phi(x_{t'}, y'))$.

A. *Multiclass Classification Metrics.*

In multiclass prediction it is very common to express the classification results as a confusion matrix or contingency table, which is not more than a two-dimensional matrix with a column and a row for each class, the row indicates the true value of the evaluated instance and each column the predicted value (See example in Fig. 1). A good classification result corresponds to large numbers within the diagonal and small numbers, ideally zero, in the items outside the diagonal [8].

| | | Assigned Class | | |
|---|---|---|---|---|
| | | A | B | C |
| Actual Class | A | 10 | 2 | 1 |
| | B | 0 | 6 | 1 |
| | C | 0 | 3 | 8 |

Fig. 1. Example of a three-class confusion matrix [2]

In the example shown in Fig. 1, with classes A, B, C, the first row of the matrix indicates that 13 objects belong to class A and 10 are correctly classified as belonging to A, two misclassified as belonging to B, and one belonging to C [2]. From the confusion matrix, the success rate can be determined:

$$Accuracy = \frac{number\ of\ correct\ classifications}{Total\ number\ of\ classifications} \qquad (4)$$

From the previous metric, the error rate can be inferred:

$$ER = 1 - Accuracy \qquad (5)$$

## III. SVM MODEL BLOCK DIAGRAM

The general components of the VSM multiclass model are those shown in Fig. 2, and the flow diagram of the developed algorithm is shown in Fig. 3.

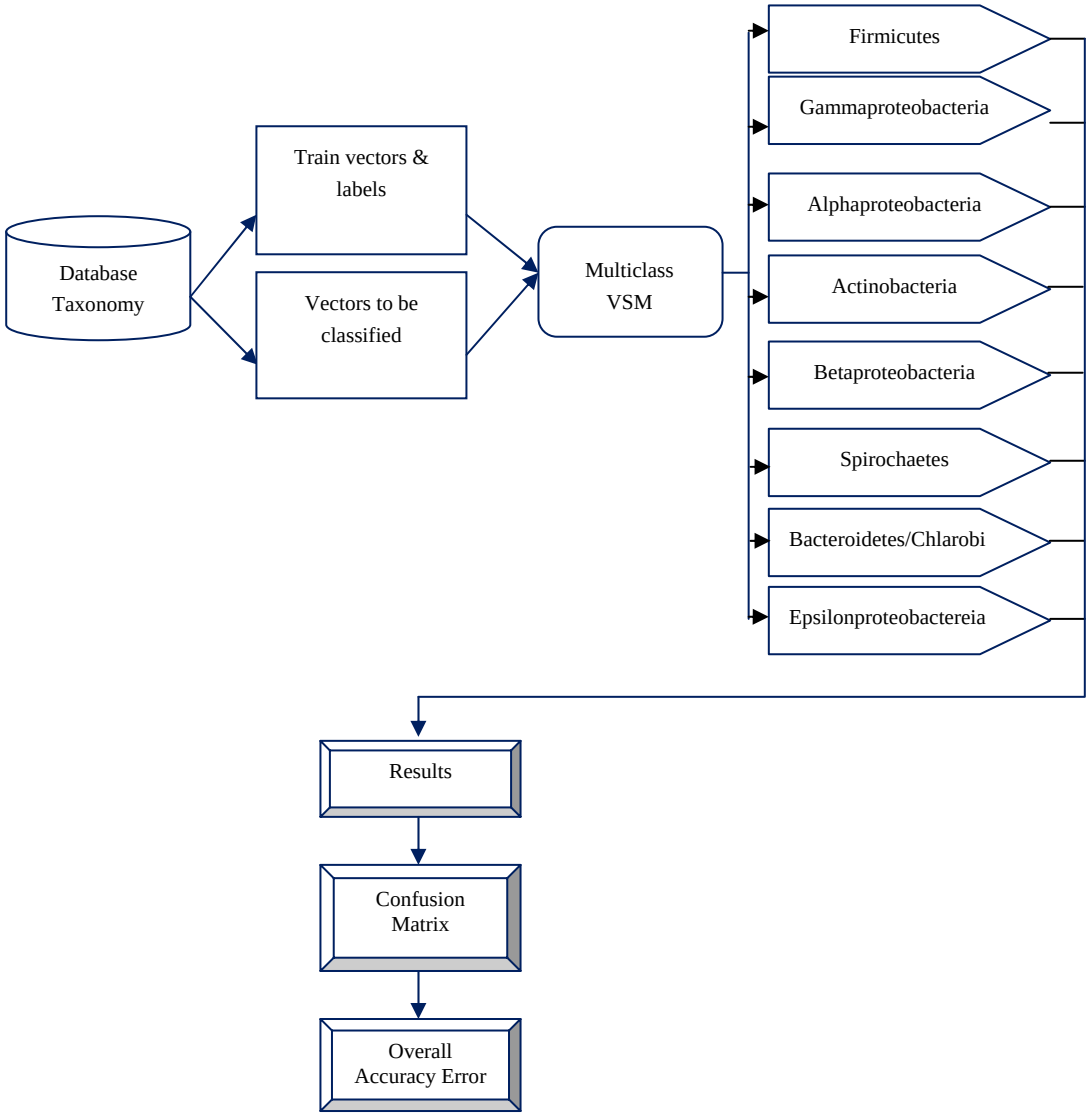


Fig. 2. Multiclass VSM model stages.

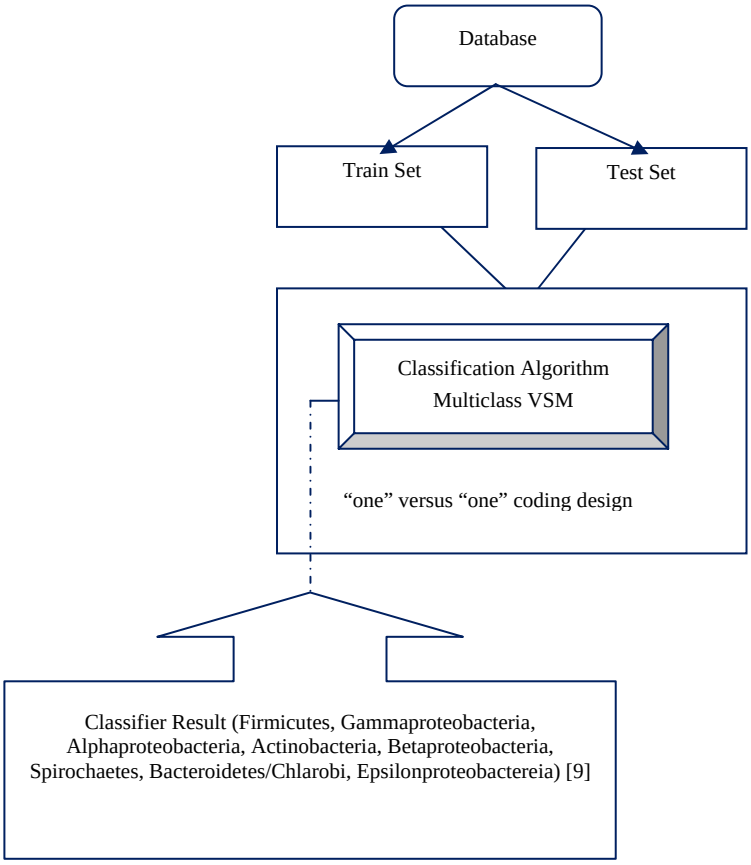


Fig. 3. Software tool flowchart of the multiclass classification system

A. *Characteristics of the Multiclass Classification Database.*

The database described in Fig. 2 and Fig. 3, which feeds the VSM, is made up of a set of data consisting of a list of 648 patterns of bacteria created randomly, with true origin labels corresponding to 8 different taxonomic families. Each pattern is made up of a sequence of 814 genes in order to attempt to emulate the database referenced in [9] and where it is important to highlight that the true labels are in fact present in that work. Figure 4 shows the distribution of samples according to taxonomic group [9].

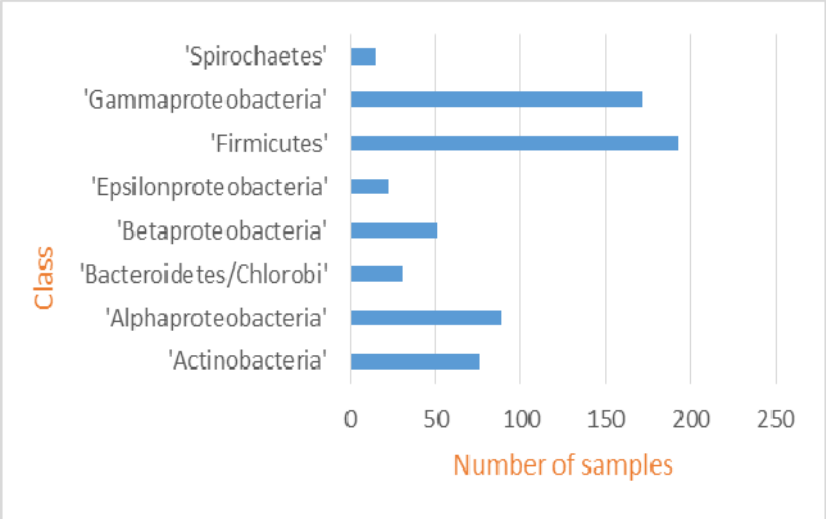


Fig. 4. Distribution of samples according to taxonomic group

## IV. CLASSIFICATION RESULTS BY TAXONOMY WITH MULTICLASS VSM

As the designed and implemented VSM addresses the problem posed by taxonomic classification of bacteria for the prediction of classes: 'Firmicutes','Gammaproteobacteria','Alphaproteobacteria','Actinobacteria', 'Betaproteobacteria','Spirochaetes', 'Bacteroidetes/Chlorobi', 'Epsilonproteobacteria', we proceed to describe the results found as detailed below.

A. *VSM Simulation Environment.*

Fig. 5 shows the classification interface (implemented in Matlab), where three areas can be seen:

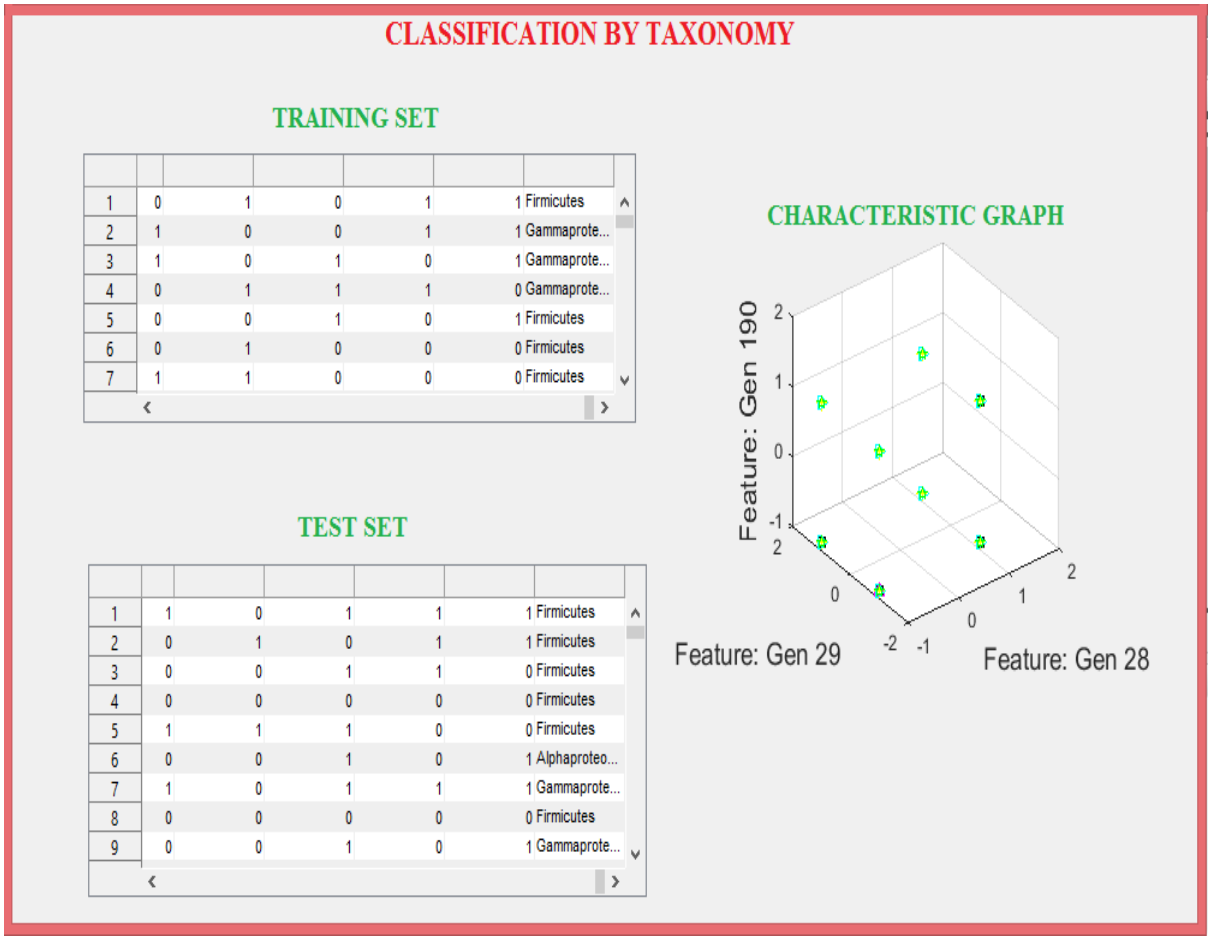


Fig. 5. Taxonomic classification software

*1) Train Set table:* Showing the patterns (samples) of bacteria that are used as input to train the VSM classifier; each pattern is composed of a sequence of orthogonal genes related to the taxonomic families of the bacteria; In addition, the current (real) labels are shown for each pattern.

2) *Test Set table:* Showing the patterns (samples) of bacteria that are used as input to the VSM classifier to be discriminated. Following each pattern is the label predicted by the classifier.

*3) Selected Characteristics graph:* Showing each one of the patterns, but in a three-dimensional space with three characteristics only.

B. *Better confusion matrices and performance for the assessed kernel functions.*

Tables I, II and III summarize the best confusion matrices and performance for the quadratic, cubic and Gaussian kernels. It should be noted that these tables show how the assignments of instances of a current class are distributed across all classes. The count within the diagonal represents the true positives for each class. The count outside the diagonal are the errors made in the assignments.

TABLE I. Confusion matrix for the quadratic kernel

| Assigned Class / Current Class | 'Firmicutes' | 'Gammaproteobacteria' | 'Alphaproteobacteria' | 'Actinobacteria' | 'Betaproteobacteria' | 'Spirochaetes' | Bacteroidetes/Chlorobi' | 'Epsilonproteobacteria' |
|---|---|---|---|---|---|---|---|---|
| **'Firmicutes'** | 162 | 31 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Gammaproteobacteria'** | 138 | 34 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Alphaproteobacteria'** | 76 | 12 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Actinobacteria'** | 64 | 12 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Betaproteobacteria'** | 41 | 10 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Spirochaetes'** | 13 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Bacteroidetes/Chlorobi'** | 27 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Epsilonproteobacteria'** | 19 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |

TABLE II. Confusion matrix for the cubic kernel

| Assigned Class / Current Class | 'Firmicutes' | 'Gammaproteobacteria' | 'Alphaproteobacteria' | 'Actinobacteria' | 'Betaproteobacteria' | 'Spirochaetes' | Bacteroidetes/Chlorobi' | 'Epsilonproteobacteria' |
|---|---|---|---|---|---|---|---|---|
| **'Firmicutes'** | 193 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Gammaproteobacteria'** | 172 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Alphaproteobacteria'** | 88 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Actinobacteria'** | 76 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Betaproteobacteria'** | 51 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Spirochaetes'** | 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Bacteroidetes/Chlorobi'** | 31 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Epsilonproteobacteria'** | 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

TABLE III. Confusion matrix for the Gaussian kernel

| Assigned Class / Current Class | Firmicutes' | 'Gammaproteobacteria' | 'Alphaproteobacteria' | 'Actinobacteria' | 'Betaproteobacteria' | 'Spirochaetes' | Bacteroidetes/Chlorobi' | 'Epsilonproteobacteria' |
|---|---|---|---|---|---|---|---|---|
| **Firmicutes'** | 193 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Gammaproteobacteria'** | 172 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Alphaproteobacteria'** | 88 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Actinobacteria'** | 76 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Betaproteobacteria'** | 51 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Spirochaetes'** | 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Bacteroidetes/Chlorobi'** | 31 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **'Epsilonproteobacteria'** | 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

C. *Analysis of Results.*

In order to assess classifier performance, the *Accuracy* (global classifier performance) and *Global error* metrics were taken into account within the cross-validation method. Each of the metrics described in Fig. 4, were calculated based on the confusion matrix of each configuration (free parameters (Table I, II and III)), which is framed within the *one versus one* ECOC classification coding design (Error Correcting Output Codes Multiclass model), which features the inherent advantages of ECOC with regard to repairing errors caused by failures generated by the finite number of train samples, number of attributes and defects in the learning process.

The best results found for the quadratic, cubic and Gaussian kernels in the *one versus one* coding design were:

$$accuracy\ quadratic\ kernel\ (C = 0.4) = \frac{162 + 34}{648} = \frac{196}{648} = 0.3025 \quad (6)$$

$$ER = 1 - 0.3025 = 0.6975$$

$$accuracy\ cubic\ kernel(C = 0.1) = \frac{193}{648} = 0.2978 \quad (7)$$

$$ER = 1 - 0.2978 = 0.7022$$

$$accuracy\ gaussian\ kernel\ (c = 0.1, \gamma = 0.1) = 193/648 = 0.2978 \quad (8)$$

$$ER = 1 - 0.3025 = 0.7022$$

The above results allow us to infer that according to the *Accuracy* metric, the best kernel for the problem posed is the quadratic kernel with a free coefficient $C = 0.4$, which has an global performance of 30.25% and an error rate of 69.75%.

## V. CONCLUSIONS

Based on the results of the research, we conclude that multiclass VSMs are not the best option to be used as an automated bacterial classification system by taxonomy when the kernel configurations are of quadratic, cubic or Gaussian type. For this reason it is worth evaluating other methodologies such as those based on neural networks, deep learning, Bayesian networks, Anfis systems in order to improve performance in taxonomic classification (i.e. seeking to reduce the ER error rate) obtained in Equations 5, 6 and 7.

## REFERENCES

[1] M. Toledo, Automatic summary and translation evaluation in the context of specialized translation. Frankfurt: Internationaler Verlag Vorbehalten, pp. 319-321, ISBN: 978-3-631-60360-4, 2010.
[2] C. Sammut, G. Webb, Encyclopedia of Machine Learning, New York, Springer, pp. 1030-1032, ISBN: 978-0-387-30164-8, 2010.
[3] A. Shigeo, Support vector machines for pattern classification (advances in computer vision and pattern recognition), Second edition, New York, Springer, pp. 21-98, ISBN: 978-1-84996-097-7, 2010.
[4] C. Cortes, V. Vapnik, Support vector networks, Machine Learning, volume 20, pp. 273–297, 1995.
[5] G. Colmenares, Artificial Intelligence, Vector Support Machine. [Online], recovered: http://www.webdelprofesor.ula.ve/economia/gcolmen/programa/economia/maquinas_vectores_soporte.pdf.
[6] D. López, E. Rivas, O. Gualdron, Characterization of primary users in cognitive radio wireless networks using Support Vector Machine, Indian Journal of Science and Technology, Volume 10, Issue 32, pp. 1-12, 2017.
[7] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun. Large margin methods for structured and interdependent output variables, Journal of Machine Learning Research, Volume 6, pp. 1453–1484, 2005.
[8] L. Ornella, Error correction codes in problems of multiclass classification of molecular marker data, Doctoral Thesis, Universidad Nacional De Rosario (UNR). Facultad de Ciencias exactas. Ingeniería y Agrimensura, P. 17, Argentina, 2010.
[9] G. Iraola, G. Vazquez, L. Spangenberg, H. Naya, Reduced set of virulence genes allows high accuracy prediction of bacterial pathogenicity in humans. PLoS ONE, Volume 7, Issue 8, 2012.