ISSN 0798 1015

REVISTA ESPACIOS

HOME Revista ESPACIOS ÍNDICES / Index A LOS AUTORES / To the AUTORS

Vol. 40 (Number 7) Year 2019. Page 29

Use of support vector machines in automated classification of bacteria by pathogenicity

Utilización de máquinas de soporte vectorial en la clasificación automatizada de bacterias por patogenicidad

HERNÁNDEZ, Leydy J. 1; LÓPEZ, Leyla C. 2; LÓPEZ, Danilo A. 3

Received: 21/08/2018 • Approved: 19/02/2019 • Published 04/03/2019

Contents

ABSTRACT:
One of the existing problems in the area of medical sciences is related to classifying pathogenic and non-pathogenic human bacteria; however with the emergence of vector support machines it is possible to optimize this task in an automated way by class separation in space called hyperplanes. In this regard, the article evaluates the performance presented when classifying bacteria using the polynomial kernel and Gaussian radial base functions. The results identified a better SVM performance when setting parameters within the biclass classification is performed using a polynomial kernel.
Keywords: Support vector machine, algorithm, biclass classification, kernel, pathogenicity

RESUMEN:
Uno de los problemas existentes en el área de las ciencias médicas se relaciona con la clasificación de bacterias patógenas y no patógenas humanas; no obstante con la aparición de las máquinas de soporte vectorial se tiene la posibilidad de optimizar dicha tarea de manera automatizada mediante la separación de clases en el espacio llamados hiperplanos. En este sentido el artículo evalúa el rendimiento presentado al clasificar bacterias utilizando las funciones kernel polinomial y base radial gaussiana. Los resultados permitieron identificar un mejor desempeño de la SVM cuando el ajuste de parámetros dentro de la clasificación biclase se realiza con el uso de un kernel polinomial.
Palabras clave: Máquinas de soporte vectorial, algoritmo, clasificación biclase, núcleo,

1. Introduction

Most bacteria are harmless to their hosts, but there are some that are pathogenic. Bacteria are grouped into taxonomic categories. Members of these categories have a sequence of orthogonal genes in some cases similar, which may indicate some information about their behavior. Although there has been major progress in understanding bacterial pathogenesis, there is still not enough information about what makes a bacterium a human pathogenic. With the advent of sequencing technologies, the number of complete bacterial genomes has

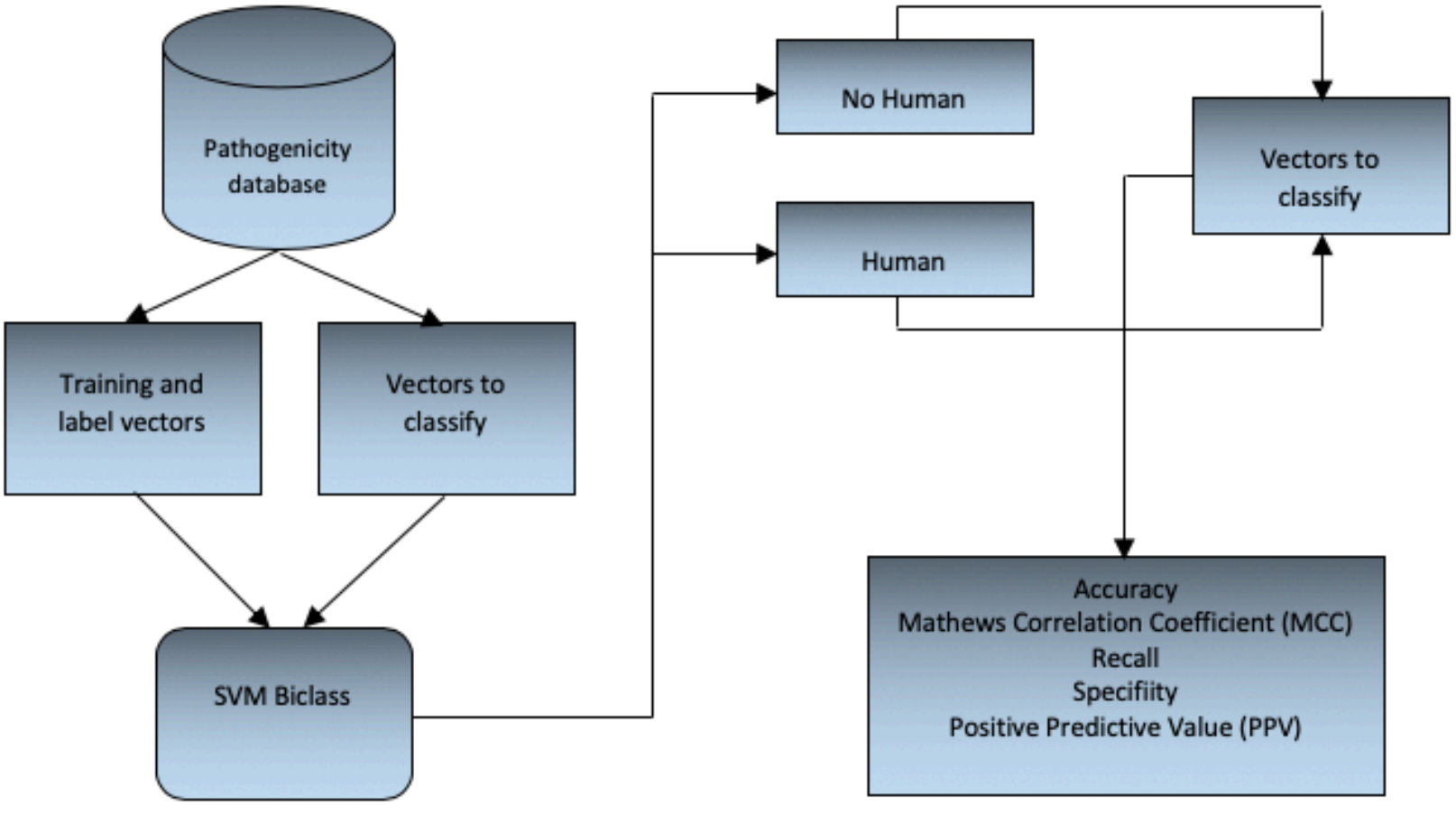
increased dramatically. This information is now available to investigate genetic characteristics that determine pathogenic phenotypes in bacteria (Iraola et al., 2012). This scenario implicitly raises the need to develop possible automated solutions to one of the problems of interest in medicine that is related to discrimination of bacteria in "human pathogenic and non-pathogenic". The above can be carried out by using classification algorithms and gene databases related to virulence, whose values depend on whether they are present or absent in the sequencing pattern. Likewise, the data processing tools, as well as the analysis of results, are very important to reach conclusions about the classification and composition of the genomic base. In this regard, results found when designing a pathogenic biclass bacteria classification system by using Support Vector Machine are socialized. (SVMs) (Rivas et al., 2017).

---

2. Methodology

To classify of bacteria into human pathogenic and non-pathogenic, the cross-validation validation method (Mathworks, 2016) has been used (Sammut and Webb, 2010), which consists of dividing the set of samples into subsets, where only a single subset is evaluated against the other training subsets during each iteration. The test subset changes in each iteration and the performance measures are calculated and finally averaged. This procedure is used for each of the configurations of the SVM coefficients and their respective kernels. Finally, the best results in performance measures will be taken into account to select the kernel and its most appropriate configuration. The general scheme of the SVM biclass model developed is shown in figure 1.

Figure 1
General flow chart scheme



It is important to emphasize that according to the kernel function that is used in the algorithm, there are free parameters that can be adjusted. By varying these parameters (known as the SVM training event) you can find the values so that the generated classifier is the best and thus adequately predict the patterns that you wish to classify (Camacho, 2013). In this sense the SVM tries to find an optimal separation hyperplane, where the margin is the highest among the groups to classify. To implement the SVM we used the Matlab programming language that has the linear kernel, quadratic, polynomial and Gaussian radial base functions implemented; and in order to determine which function had the best performance (from the metrics described in figure 1), each of these kernels was

evaluated by adjusting the respective free parameters. To adjust the parameters we used the grid-search method (Chih-Jen, 2003) which consists in the use of a mesh where the values of the SVM parameters are tested and the most accurate configuration obtained when using the cross-validation validation method is selected.

2.1. Performance evaluation variables for SVM classification.

Metrics taken into account for the classification tests in the present project are: Accuracy, Matthews correlation coefficient (MCC), Recall, Specifity, Positive Predictive Value (PPV). These measures are built based on the following values (Classeval-wordpress, 2017):

Accuracy. It is the proportion of true results (true positives and true negatives) among the total number of cases evaluated:

$$\frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

MCC. It is a measure of the quality of binary classifications it is mainly used in binary classifiers:

$$\frac{TP.TN - FP.FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2)$$

Recall. It is the ability of the classifier to find all the positive samples:

$$\frac{TP}{TP + FN} \quad (3)$$

Said variable was calculated as the number of human pathogenic samples correctly classified divided by the number of true human pathogenic samples.

Specifity. It is the fraction of negative samples correctly predicted by a model:

$$\frac{TN}{(TN + FP)} \quad (4)$$

In this project it was calculated as the number of non-pathogenic human samples correctly classified, divided by the number of true non-pathogenic human samples.

PPV. It is the proportion of true positives for all instances predicted as positive:

$$\frac{TP}{(TP + FP)} \quad (5)$$

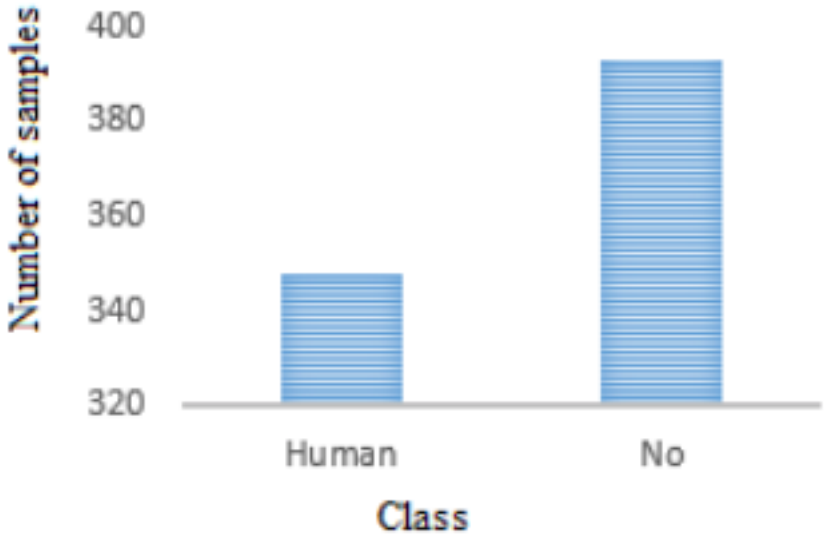
In this project it was calculated as the number of human pathogenic samples correctly classified, divided by the number of samples classified as human pathogens. It should be noted that for each of the equations, the "correct classifications" correspond to: TP = true positives, TN = True negatives; whereas the "incorrect classifications" are of type: FP = False positives, FN = False negatives.

2.2. Characteristics of the biclass classification database.

It consists of a set of data comprising a list of 741 bacterial patterns with their corresponding true labels. Each pattern consists of a sequence of 120 genes, which were reduced from a larger set of 814 genes in the work developed in Iraola et al. (2012) to increase the efficiency of both classification performances, as in computer processing. This

data set was obtained from the database that accompanies the software Bacfier_v1_0.zip of free distribution, developed by the authors of the work and available in Iraola et al. (2012). Figure 2 illustrates the distribution of samples.

Figure 2
Distribution of samples according to pathogenicity



3. Results

Tables 1 to 3 show the results of the adjustment meshes of free parameters (*c*, *d*) found with the polynomial kernel for human pathogenicity classification, and in Tables 4 to 6 the results for the Gaussian radial base (*c*, *γ*) kernel function, in accordance with what is established in Camacho (2013). Due to lack of space the results found with the linear and quadratic kernel are not included.

In each case, in order to determine the best performance for the metrics (Accuracy MMC, Recall, Specifity and PPV) the respective free parameters were adjusted; and to adjust these parameters the "grid search" method was used, which consists in the use of a mesh where the values of the SVM parameters are being tested and the configuration with the best precision obtained when using the cross-validation method is selected.

Table 1
Mesh to set the parameters of the *Accuracy* and *MCC* variables for the polynomial kernel function.

c / d	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
1	0.939	0.949	0.947	0.941	0.923	0.946	0.927	0.912	**0.950**	0.915
2	0.907	0.906	0.923	0.916	0.918	0.927	0.926	0.914	0.930	0.924
3	0.877	0.922	0.915	0.939	0.924	0.922	0.880	0.904	0.911	0.906
4	0.765	0.765	0.754	0.757	0.769	0.758	0.779	0.761	0.754	0.754
5	0.541	0.545	0.540	0.547	0.543	0.536	0.543	0.548	0.540	0.541

c / d	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
1	**0.442**	0.436	0.432	0.437	0.429	0.437	0.430	0.433	0.435	0.428
2	0.399	0.416	0.411	0.404	0.414	0.410	0.410	0.397	0.413	0.409
3	0.427	0.428	0.424	0.436	0.433	0.428	0.426	0.418	0.424	0.428
4	0.293	0.286	0.272	0.289	0.283	0.293	0.306	0.301	0.292	0.281
5	0.055	0.066	0.045	0.074	0.060	0.025	0.058	0.078	0.047	0.053

Table 2
Mesh to set the *Recall* and *Specifity* variable parameters for the polynomial Kernel function.

C / d	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
1	0.969	0.957	0.952	0.962	0.959	0.959	0.959	0.969	0.954	0.962
2	0.926	0.952	0.934	0.929	0.941	0.931	0.931	0.919	0.934	0.931
3	**0.980**	0.959	0.957	0.962	0.964	0.959	0.977	0.954	0.959	0.967
4	0.875	0.863	0.850	0.878	0.852	0.883	0.883	0.893	0.885	0.868
5	0.977	0.972	0.972	0.975	0.977	0.967	0.975	0.975	0.975	0.975

C / d	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
1	0.905	0.940	0.943	0.917	0.882	0.931	0.891	0.848	**0.945**	0.862
2	0.885	0.853	0.911	0.902	0.891	0.922	0.920	0.908	0.925	0.917
3	0.761	0.879	0.868	0.914	0.879	0.879	0.770	0.848	0.856	0.836
4	0.641	0.655	0.647	0.621	0.675	0.618	0.661	0.612	0.606	0.626
5	0.049	0.063	0.052	0.063	0.052	0.049	0.055	0.066	0.049	0.052

Table3
Mesh for setting *PPV* variable parameters
for the polynomial kernel Function

C / d	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
1	0.920	0.947	0.949	0.929	0.902	0.940	0.908	0.878	**0.952**	0.887
2	0.901	0.880	0.922	0.915	0.907	0.931	0.929	0.919	0.934	0.927
3	0.823	0.900	0.891	0.926	0.900	0.900	0.828	0.876	0.883	0.870
4	0.733	0.739	0.731	0.723	0.748	0.723	0.746	0.722	0.718	0.724
5	0.537	0.540	0.537	0.540	0.538	0.534	0.538	0.541	0.536	0.537

At this point it is important to note that although the polynomial kernel function has more than two free parameters, it was decided to use only parameters c and d, due to the difficulties that would arise to form the mesh (Camacho, 2013). It should be noted that for the different tables the regularization parameter c, is the variable that establishes the relationship between the training error and the complexity of the model (the higher the value of c, the more complex the model and the lower the training error, its value typically varies between 0 and 1); the parameter d, is related to the degree of the polynomial and as its value increases the classification surface becomes more complex and its value in the tests was limited to a maximum of 5 due to the restrictions in the computation processing; and the metric γ, is responsible for controlling the shape of the separation hyperplane.

Table 4
Grid to set the *Accuracy* and *MCC* variable parameters
for the Gaussian radial basis kernel function.

C / γ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,1	0.610	0.663	0.686	0.703	0.703	0.767	0.764	0.767	0.769	0.767
0,2	0.603	0.649	0.686	0.703	0.694	0.771	0.767	0.772	0.773	0.772
0,3	0.618	0.646	0.690	0.703	0.698	0.772	0.772	0.771	0.767	0.764
0,4	0.607	0.660	0.691	0.706	0.700	0.765	0.769	0.767	0.764	0.767
0,5	0.610	0.660	0.690	0.703	0.696	0.769	0.769	0.772	0.761	0.768
0,6	0.617	0.650	0.695	0.703	0.698	0.764	0.767	0.764	0.769	0.771
0,7	0.609	0.661	0.690	0.698	0.706	0.773	0.769	0.769	0.761	0.768
0,8	0.598	0.664	0.687	0.706	0.718	0.772	0.769	0.767	0.767	0.767
0,9	0.622	0.655	0.683	0.703	0.726	0.769	0.769	0.767	0.761	0.769
1	0.611	0.655	0.690	0.700	0.729	0.768	0.771	0.761	0.761	**0.776**

C / γ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,1	0.255	0.316	0.336	0.350	0.350	0.390	0.388	0.390	0.391	0.390
0,2	0.246	0.302	0.336	0.350	0.343	0.395	0.390	0.393	0.396	0.393
0,3	0.266	0.300	0.340	0.350	0.346	0.393	0.393	0.395	0.390	0.388
0,4	0.252	0.313	0.341	0.353	0.348	0.392	0.391	0.390	0.388	0.390
0,5	0.255	0.313	0.340	0.350	0.345	0.391	0.391	0.393	0.389	0.393
0,6	0.265	0.304	0.344	0.350	0.346	0.388	0.390	0.388	0.391	0.395
0,7	0.254	0.315	0.340	0.346	0.353	0.396	0.391	0.391	0.387	0.393
0,8	0.238	0.317	0.338	0.353	0.361	0.393	0.391	0.390	0.390	0.390
0,9	0.272	0.308	0.334	0.350	0.367	0.391	0.391	0.390	0.387	0.391
1	0.257	0.308	0.340	0.348	0.369	0.391	0.395	0.387	0.387	**0.398**

Table 5
Grid to set the *Recall* and *Specifity* variable parameters
for the Gaussian radial basis kernel function

C / γ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,1	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.997	0.997	0.997	0.997	0.997
0,2	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.997	0.997	**1.000**	0.997
0,3	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.997	0.997	**1.000**	0.997	0.997
0,4	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.997	0.997	0.997	0.997
0,5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.997	0.997	0.997	**1.000**	**1.000**
0,6	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.997	0.997	0.997	0.997	**1.000**
0,7	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.997	0.997	0.997	**1.000**
0,8	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.997	0.997	0.997	0.997	0.997
0,9	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.997	0.997	0.997	0.997	0.997
1	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.997	**1.000**	0.997	0.997	**1.000**

C / γ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,1	0.170	0.282	0.330	0.368	0.368	0.506	0.500	0.506	0.511	0.506
0,2	0.155	0.253	0.330	0.368	0.348	0.511	0.506	0.517	0.517	0.517
0,3	0.187	0.247	0.339	0.368	0.356	0.517	0.517	0.511	0.506	0.500
0,4	0.164	0.276	0.342	0.374	0.362	0.500	0.511	0.506	0.500	0.506
0,5	0.170	0.276	0.339	0.368	0.353	0.511	0.511	0.517	0.491	0.506
0,6	0.184	0.256	0.351	0.368	0.356	0.500	0.506	0.500	0.511	0.511
0,7	0.167	0.279	0.339	0.356	0.374	0.517	0.511	0.511	0.494	0.506
0,8	0.144	0.284	0.333	0.374	0.399	0.517	0.511	0.506	0.506	0.506
0,9	0.195	0.264	0.325	0.368	0.417	0.511	0.511	0.506	0.494	0.511
1	0.172	0.264	0.339	0.362	0.422	0.509	0.511	0.494	0.494	**0.523**

Table 6
Grid for setting *PPV* variable parameters for
Gaussian radial basis kernel function

C / γ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,05	0.576	0.611	0.628	0.641	0.641	0.695	0.693	0.695	0.698	0.695
0,1	0.572	0.602	0.628	0.641	0.634	0.698	0.695	0.700	0.701	0.700
0,15	0.581	0.600	0.631	0.641	0.637	0.700	0.700	0.698	0.695	0.693
0,2	0.575	0.609	0.632	0.643	0.639	0.693	0.698	0.695	0.693	0.695
0,25	0.576	0.609	0.631	0.641	0.636	0.698	0.698	0.700	0.689	0.696
0,3	0.581	0.603	0.635	0.641	0.637	0.693	0.695	0.693	0.698	0.698
0,35	0.575	0.610	0.631	0.637	0.643	0.701	0.698	0.698	0.690	0.696
0,4	0.569	0.612	0.629	0.643	0.653	0.700	0.698	0.695	0.695	0.695
0,45	0.584	0.606	0.626	0.641	0.659	0.698	0.698	0.695	0.690	0.698
0,5	0.577	0.606	0.631	0.639	0.662	0.696	0.698	0.690	0.690	**0.703**

3.1. Analysis of results and discussion

In the SVM biclass corresponding to the problem of classifying patterns between the human pathogenic class and non-human pathogenic class, a series of performance measures that have been calculated for each of the configurations of parameters free of SVM and the kernel were taken into account (as seen in Tables 1 to 6), which allowed to determine the best configuration in terms of performance in the classification (López et al., 2017).

Initially when analyzing the conformation of the data set, it can be observed that comparing the number of patterns of the human pathogenic class (348 patterns) with those of the non-human pathogenic class (393 patterns), there is no problem of significant sample imbalance, that does occur in similar approaches of biclass classification as in the case of cancer detection addressed by Laza and Pavón (2009), which has few samples of sick patient class as opposed to the large number of samples of the healthy patient class, and in which the idea is expressed that the global error tries to be reduced by the classifier without taking into account the data distribution. Likewise, the number of samples of the training set is large (741 * 0.9 = 667 patterns for training), which translates into a low variance, and that the number of characteristics is also large, ensuring that there is sufficient discrimination information and low oscillation in the algorithm. By having a low variance and oscillation, a good performance is obtained in the test set. It should also be noted that although the number of features is relatively large, the number of samples is higher, which largely prevents over fitting. Measures or metrics taken into account to measure classifier performance with the "cross-validation" method described above in Figure 1.

In this regard, in the first instance observing the *Accuracy* measure, shown in the left part of Tables 1 and 4, and corresponding to the classifier global performance measurement, it can be seen that the best configuration corresponds to the polynomial kernel with free parameters and .

The second measure observed globally regarding the classification is the metric called *MCC* (located on the right side of tables 1 and 4) which relates to the binary classifications quality measurement; it shows that the highest values are in the polynomial kernel, largely corroborating the results obtained with the *Accuracy* variable.

Comparing the measurements obtained for *Recall* and *Specifity* of Tables 2 and 5 metric, are measures focused on one of the classes; for the binary classification problem raised, the human pathogenic class corresponds to the *Recall* variable and the non-pathogenic human class corresponds to the *Specifity* metric. The best reported *Recall* values are those who correspond to the Gaussian kernel (100% *Recall*), but all of the other performance measure values are not the best, so it can be concluded that the kernel that performs best is the polynomial with a 98% performance for the free parameters configuration (see left part of table 2). Regarding the *Specifity* variable it can be seen that the best configuration is also

the polynomial with a performance of 94.5%.

Finally, the *PPV* metric is taken into account in the analysis (tables 3 and 6), also called precision, whose best performance is the polynomial configuration with 95.2% performance. According to Bridge Derek, the *Recall* and *PPV* measures are focused on the positive class (usually), that for the approach of this research is the human pathogenic class, which is special with respect to the nature of the problem. Similarly, sensitivity is the proportion of examples of the positive kind which are correctly classified, and the positive predictive value (*PPV*) is the proportion of examples of the positive class assigned to this class which are properly classified. According to (Ng, 2015), the two measures are robust in terms of influences that can be derived from the imbalance of classes should this occur in a significant way, since these measures give a better understanding of the predictor algorithm performance, because if in some case the situation arises that the classifier predicts few or no true positives, the sensitivity would give a low value, which would indicate a bad classification on the class and therefore a low performance of the classified. The results described above indicate that in terms of evaluating the classification performance, the best kernel is the polynomial, and that the best configuration of parameters within it is y , since most of the best metrics are in that configuration. Regarding the computational cost, it was established that the polynomial kernel is not the most adequate, since it needs a greater number of iterations than the Gaussian kernel to carry out the classification.

4. Conclusions

Taking as a reference the results found in the investigation, it was concluded that biclass SVMs are a good bacteria system classifying by pathogenicity when the kernel configuration is polynomial; however, it has the disadvantage of requiring higher data processing capacity compared to using a Gaussian radial basis kernel. For this reason it is interesting to continue evaluating other types of models or methodologies such as those based on deep learning, Bayesian networks, Anfis systems in order to maintain or improve bacteria classification performance but seeking to reduce the number of iterations required for data processing.

Bibliographic references

Camacho, F (2013). Classification system of antibacterial peptides using vector support machines. Thesis, Universidad Industrial de Santander. Faculty of Engineering and Information Systems, Bucaramanga.

Chih-Jen, L (2003). A practical guide to support vector classification, Department of Computer Science, National Taiwan University. [On line], recovered from: http://ntur.lib.ntu.edu.tw/bitstream/246246/20060927122852476378/1/freiburg.pdf

Classeval-wordpress (2017). Basic evaluation measures from confusion matrix. [On line], recovered from: https://classeval.wordpress.com/introduction/basic-evaluation-measures/

Iraola, G., Vazquez, G., Spangenberg, L., Naya, H (2012). Reduced Set of Virulence Genes Allows High Accuracy Prediction of Bacterial Pathogenicity in Humans. PLoS ONE, Volume 7, No. 8. https://doi.org/10.1371/journal.pone.0042144.

Laza, R y Pavón R (2009). Bayesian classifier of MedLine documents from unbalanced data. Master's Thesis, Universidad de Vigo, Spain.

Rivas, T., López, D., Gualdron, E (2017). Characterization of primary users in cognitive radio wireless networks using Support Vector Machine, Indian Journal of Science and Technology, Volume 10, No. 32, pp. 1-12. https://doi.org/10.17485/ijst/2017/v10i32/93796

López, D., Hernández, L., Rivas, T (2017). SVM and ANFIS as channel selection models for the spectrum decision stage in cognitive radio networks, Contemporary Engineering Sciences, Volume 10, No. 10, pp. 475-502. https://doi.org/10.12988/ces.2017.7438.

Mathworks, Matlab (2016). Cross-validated support vector machine regression model. [On line], recovered from: https://www.mathworks.com/help/stats/regressionpartitionedsvm-class.html?searchHighlight=crossvalidatedmodel&s_tid=doc_srchtitle

Ng, A (2015). Machine learning, Stanford University, Video lectures. [On line], recovered

from: https://es.coursera.org/learn/machine-learning

Sammut, C y Webb, G (2010). Encyclopedia of Machine Learning. Ed: Springer, pp. 1031, New York. ISBN 978-0-387-30164-8.

---

1. Estudiante de Doctorado en Ingeniería, Docente-Investigador de la Unidad de Ingenierías, en el Programa de Tecnología en Logística Empresarial. Corporación Universitaria Minuto de Dios - UNIMINUTO. leydy.hernandez-v@uniminuto.edu.co

2. Magister en Psicología, Profesora en la Facultad de Ciencias Humanas y Educativas en el Programa de Psicología. Universidad de Boyacá. leycarlopez@uniboyaca.edu.co

3. Doctor en Ingeniería. Profesor en la Facultad de Ingenierías en el Programa de Ingeniería Eléctrica. Universidad Distrital Francisco José de Caldas. dalopezs@udistrital.edu.co

---

Revista ESPACIOS. ISSN 0798 1015
Vol. 40 (Nº 07) Year 2019

[Index]