**Modelo predictivo para el rendimiento de cultivos de cacao en Santander basado en herramientas de aprendizaje automático supervisado**

PhD. Henry Lamos Díaz1, David E. Puentes Garzón2, Andrea C. Gamboa Ariza3, Paula A. Cáceres Ortiz4

Escuela de Estudios Industriales y Empresariales -EEIE

Grupo de Investigación Ópalo

Universidad Industrial de Santander

hlamos@uis.edu.co1, dpuentesgarzon@gmail.com2, andrea1496gamboa@gmail.com3, paulacaceresor@gmail.com4

Abstract

En el presente artículo se exponen dos modelos para la predicción de rendimientos de cacao en Santander: Modelo Lineal Generalizado (GLM) y Máquinas de Soporte Vectorial (SVM). Los GLM son una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores (binomiales, Poisson, gamma, etc.) y varianzas no constantes. Estos modelos son una alternativa a la transformación de la variable respuesta y a la falta de normalidad en el modelo, mientras que una Máquina de Soporte Vectorial (SVM) aprende la superficie decisión de dos clases distintas de los puntos de entrada. Diámetro del tronco, Fósforo (P), Magnesio (Mg), %Arena, %Hum/Grav, Radiación, Temperatura, Humedad y Lluvias acumuladas fueron seleccionadas como variables de entrada para los modelos. Así mismo, estos modelos se evaluaron con diferentes métricas de ajuste obtenidas a partir de validación cruzada, obteniendo resultados de R^2:(0.1534) RMSE: (144.1014) MAE: (854.4128) para SVM Y R^2: (0.2336) RMSE: (143.9239) MAE: (928.545) para GLM, resultados importantes al momento de predecir rendimientos de cacao.

1. Introducción

La agricultura es una de las actividades de mayor contribución al crecimiento económico de la población en Colombia. Particularmente, el cultivo de cacao contribuye en gran medida a dicho crecimiento, debido a que fue el cultivo que más creció porcentualmente en producción; de 2016 a 2017 se produjo un incremento en la producción del 6.6% el cual creó un récord para el país en este sector [1].

Del mismo modo, el sector cacaotero está catalogado como uno de los sectores estratégicos en el departamento de Santander con un porcentaje de participación en el área nacional sembrada del 24% y una producción de aproximadamente 26.431,64 Ton/Año [2].

En aras de contribuir con uno de los objetivos del Plan de Desarrollo Departamental, el cual busca “fortalecer la agricultura familiar de tal forma que se garantice la seguridad alimentaria”, se origina el presente trabajo en el que por medio de las metodologías de aprendizaje automático: Modelo lineal Generalizado y Máquinas de Soporte Vectorial, se pretende predecir el rendimiento de cultivos de cacao.

Las herramientas de aprendizaje automático han sido aplicadas para la predicción de diferentes cultivos agrícolas en trabajos como los de Chen et al. (2016) & Chattopadhyay & Mitra (2018), en los que se utilizó gran cantidad de datos de entrada para predecir la variable respuesta para cultivos de arroz y granos respectivamente.

Tanto Máquinas de Soporte Vectorial como Modelo Lineal Generalizado, han sido técnicas de Aprendizaje Automático que han ayudado a encontrar influencias, dentro de los cultivos, de factores climáticos, morfológicos y de suelo, y con ello, han facilitado a los diferentes actores a formular estrategias adecuadas para hacerle frente a las variables con mayor influencia dentro de los cultivos.

En ese orden de ideas, predecir el rendimiento para los cultivos alimenticios y determinar variables influyentes en ellos, es de gran ayuda para que agricultores, empresas y gobierno puedan comprender las condiciones mas influyentes y en consecuencia, mejorar la adaptabilidad de los cultivos tomando decisiones acertadas anticipadamente con el objetivo de obtener mayor rendimiento en los cultivos.

Protein sequences can be mapped into feature vectors by extracting three types of attributes: Primary structure statis­ tics, secondary structure statistics and physical-chemical features. Then, a decision function can be learned by using pattern recognition tools such as support vector machines (SVM) [2] which have been successfully applied to many

The process of deriving a decision rule from input at­ tributes is called "training". Basically, a large number of la­ beled data are presented to the algorithm with his attributes as a way to teach the variation pattern. This process may affected by the number of objects belonging to one class or another, which is also known as class distribution. If the number of objects is the same in all classes the class distribution is said to be balanced, in other case, the class distribution imbalanced [4, 5, 6].

The success of most classifier algorithms, which assume a relatively balanced distribution, decreases when they are used to the problem of learning from imbalanced datasets, as it is the case of protein function prediction [7]. For this reason, the problem of learning from imbalanced data has been intensively researched in the last decade and several methods proposed to discuss it. [8] developed an over­ sampling method called synthetic minority over-sampling technique (SMOTE) that generates new artificial examples for the positive class or minority class by interpolating among several minority class examples that lie together. This method have been already succesfully used for protein function prediction [9], but the inclusion of synthetic sam­ ples produces a high computational cost that is not desirable in this kind of applications.

Sub-sampling is a method that eliminates objects of the greater class or negative class, trying to reduce the class imbalance. A few methods using sub-sampling for pro­ tein function classification have been proposed in the recent years [10], but this method generally produces an undesir­ able loss of information, because the removed objects are real data [11, 4, 5].

A third option relies on cost-sensitive methods. This kind of methods attempts to minimize costs associated to their decisions rather than simply reaching high precision. In the case of SVM, several authors have proposed al­ gorithm variations using different penalty parameters in

978-1-4673-9461-1/15/$3l.00 ©2015 IEEE

the SVM formulation [12]. This approach is known as weighted SVM (WSVM) and does not imply a resampling of the data, which makes it a good candidate for the task of protein function prediction.

This paper presents a comparison of three strategies for managing the imbalance problem: undersampling, SMOTE and WSVM, in order to analyse their performance in the specific task of protein function prediction. Comparisons are made in terms of classification performance (measuring specificity, sensitivity and geometric mean) and computa­ tional time.

2. Materials and methods

The overall methodology used in this work is depicted in figure 1. Each stage of such methodology will be detailed in the next sections.

METHODOLOGY

�s

Figure 1. Proposed methodology

2.1. Molecular function data

Database is made up of fourteen different classification problems, belonging to the molecular function ontology of Gene Ontology (GO). This database is a subset of the data used in [9] and includes 2544 proteins associated to the Em­ bryophyta taxonomy of the Uniprot database [13] with an­ notation in the GO project [1]. The dataset does not contain any protein sequences with a sequence identity higher than

30%. Table 1 shows the number of sequences associated to each problem.

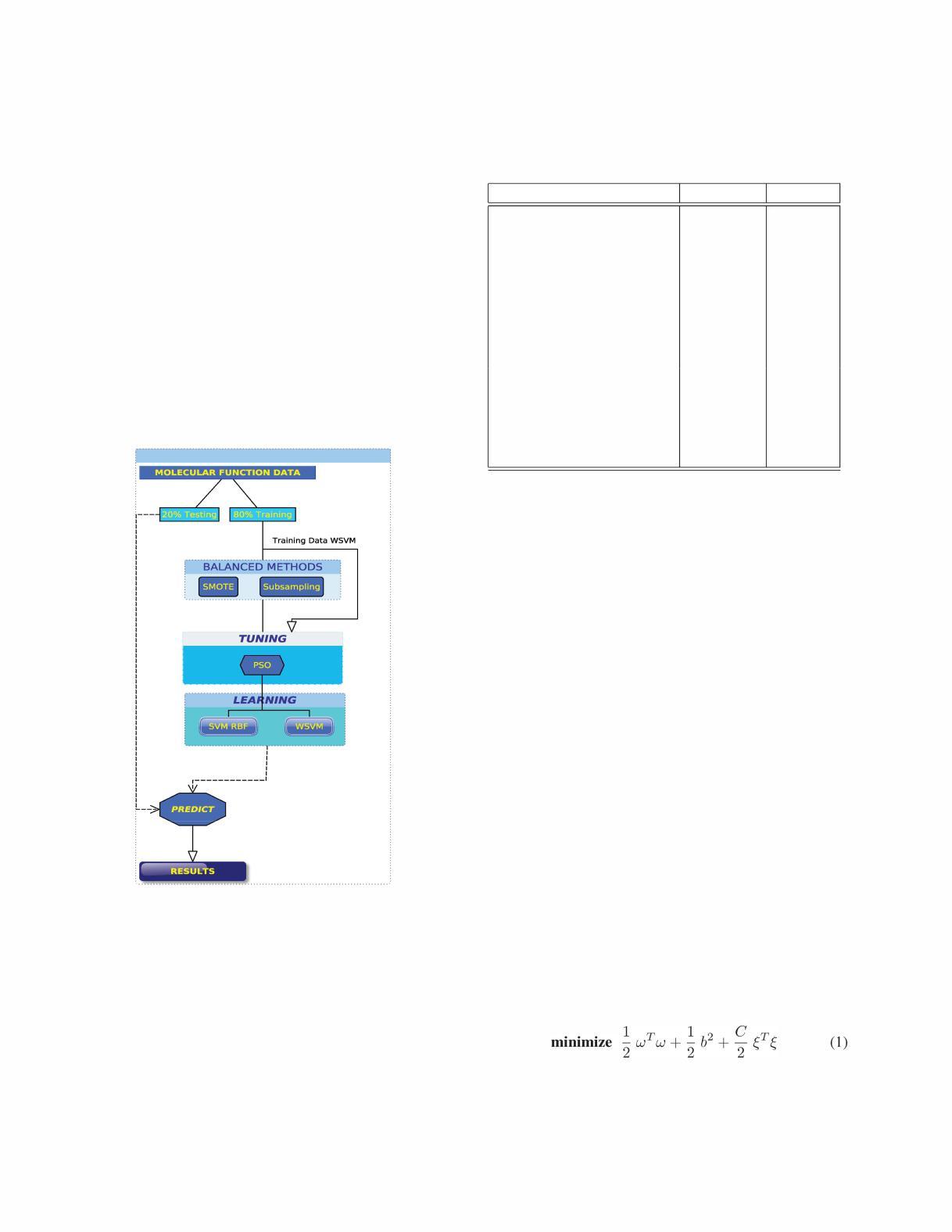


Table 1. Number of sequences for each GO term

Molecular Function I Acronym I Samples I

|  |  |  |
| --- | --- | --- |
| Nucleotide binding | Ntbind | 47 |
| Molecular function | MF\* | 268 |
| DNA binding | DnaBind | 107 |
| Transcription factor activity | TranscFact | 307 |
| RNA binding | RnaBind | 43 |
| Catalytic activity | Catal | 334 |
| Receptor binding | RecBind | 38 |
| Transporter activity | Transp | 125 |
| Binding | Bind | 173 |
| Protein binding | ProtBind | 630 |
| Kinase activity | Kinase | 68 |
| Transferase activity | Transf | 173 |
| Hydrolase activity | Hydrol | 190 |
| Enzyme regulator activity | EnzReg | 41 |

2.2. Sampling methods

2.2.1 Synthetic Minority Oversampling Technique (SMOTE)

This approach is based on generating new instances by in­ terpolation. For each positive instance, its nearest positive neighbors are identified and new positive instances are cre­ ated and placed randomly in between the instance and its neighbors [8]. Since this technique creates new positive in­ stances, we found this technique to be more useful for SVM than simple oversampling.

2.2.2 Subsampling

The subsampling method employed uses The K-Means al­ gorithm in order to find narrowly most reprsentative sam­ ples in the mjority class, by minimizing the sum of the squared distance between each data point and its proximal cluster center. The number of clusters is fixed to the num­ ber of samples in the minority class and thus, cluster centers provide a balanced set.

2.3. Weighted SVM (WSVM)

The WSVM applies quadratic cost function on the slack variable and the bias term is also included in the object func­ tion. WSVM is formulated as

|  |  |  |
| --- | --- | --- |
| s. t . | { | Y(Xw + be) - e + � ;:::0, |
|  | � ;:::0 |
|  |  |

2

(gm)

To deal with the imbalanced dataset, a weight Wi is assigned to each data point to and the weighted WSVM is formulated as

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| minimize -l | | wT w + -I | b2 + -e | �TW� | (2) |
|  | 2 | 2 | 2 |  |  |
| s.t. { | Y (X w + be) - e + � � 0, | | | |  |
| ��O | |  |  |  |

2.4. Validation

Three performance measures are used to analyse the gen­ eralization capability of the predictor: sensitivity (sn ) de­ scribes the capacity of the algorithm to recognize as posi­ tives the sequences that are indeed associated to a given sub­ cellular component; specificiy (sp ) describes how the algo­ rithm is able to reject sequences that are not associated to

the sub-cellular component; and the geometric mean

between those measures as a global performance measure [6].

S TP TN

n = TP +FN Sp = TN +FP

TPTN

gm = (T P + F N) (TN + F P)

where TP, nTN, FP and FN are true positive, true nega­ tive, false positive, and false negative, respectively.

2.5. Parameters tuning

Tuning is carried out by a Particle Swarm Optimization

(PSO) [14] which explores search space generated by all the possible pairs of values that can be cOlmnit to the trade-off

constant of the SVM (C) and the dispersion parameter of the gaussian kernel (0-). To this end, a new partition on the training set is done following a cross-validation of ten folds, in order to avoid over-training of the models.

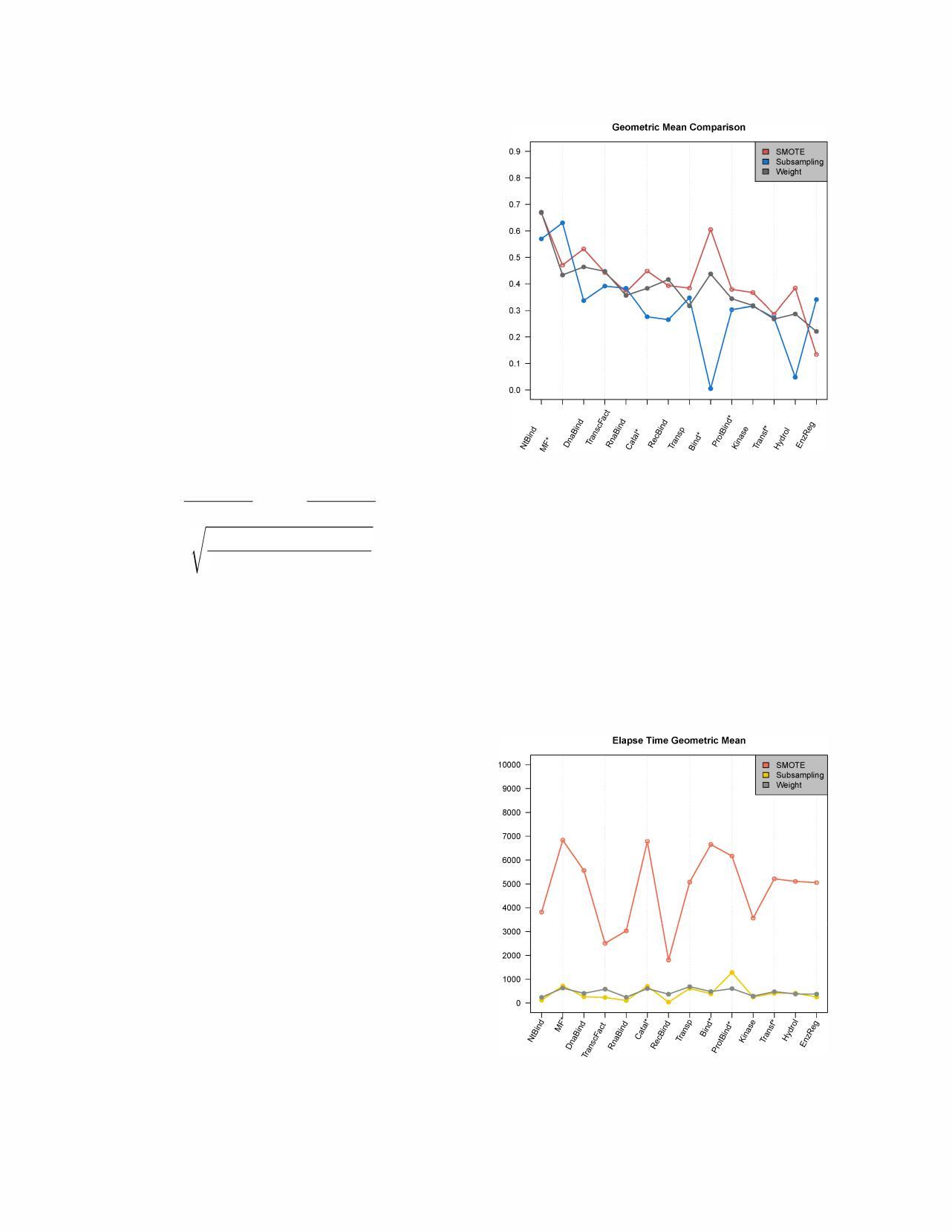
3. Results and discussion

As can be observed on Table 2, subsampling provides better sensibilities for 12 out of the 14 molecular functions considered in the database, which means that the classifier is able to succesfully identify the samples of the positive class. On the other hand, SMOTE achieved the highest specifici­ ties, demonstraiting that this method is better able to cor­ rectly exclude samples of the negative class.

As a global performance measure, geometric mean be­ tween both specificity and sensitivity was computed for all classes and the results are depicted on figure 3. This figure shows that SMOTE and WSVM present comparable per­ formances in most of the problems, while subsampling is clearly under those two methods, with a few exceptions.

Finally, in order to compare the computational complex­ ity of the methods, figure [?] depicts times (in seconds)

Figure 2. Decreased Geometries means to: SMOTE,Subsampling and Weighted Lagrangian SYM.

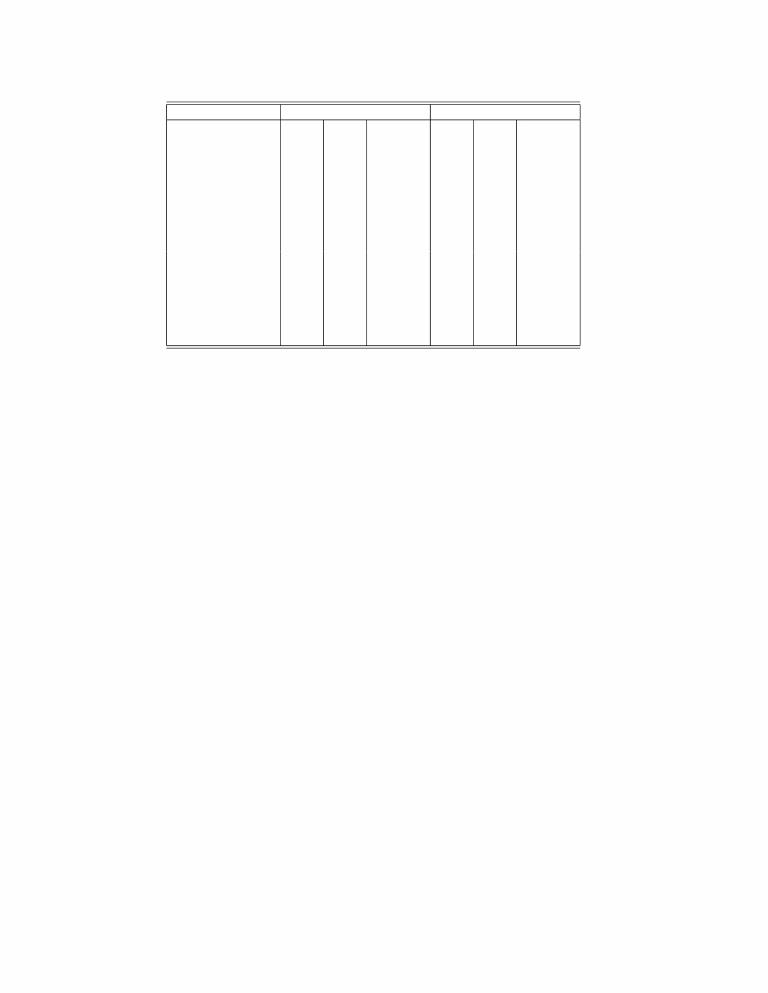


spent by each one of the methodologies as explained on figure l. It is possible to see that SMOTE takes the high­ est time, presumably due to the increase in the size of the databases induced by the incorporation of synthetic sam­ ples.

SMOTE balancing method highly efficient but has an el­ evate computational cost involved in less efficiency; in turn, WSVM maintains a very high stability between classifica­ tion and computing efficiently consumption.

Figure 3. Geometric mean and standard deviation of times in all issues

3

Table 2. Results comparisson SMOTE,Subsampling and WLSVM

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Molecular Function |  | Sensitivity | |  | Specificity | |
|  | SM | SS | WSVM | SM | SS | WSVM |
| Ntbind | 0.234 | 0.297 | 0.234 | 0.766 | 0.705 | 0.793 |
| MF | 0.395 | 0.723 | 0.500 | 0.843 | 0.620 | 0.769 |
| DnaBind | 0.607 | 0.906 | 0.654 | 0.737 | 0.207 | 0.677 |
| TranscFact | 0.726 | 0.850 | 0.791 | 0.911 | 0.657 | 0.744 |
| RnaBind | 0.418 | 0.534 | 0.604 | 0.746 | 0.722 | 0.704 |
| Catal | 0.682 | 0.931 | 0.580 | 0.605 | 0.142 | 0.620 |
| RecBind | 0.263 | 0.842 | 0.605 | 0.998 | 0.906 | 0.987 |
| Transp | 0.648 | 0.960 | 0.600 | 0.930 | 0.175 | 0.903 |
| Bind | 0.479 | 0.641 | 0.341 | 0.708 | 0.548 | 0.808 |
| ProtBind | 0.603 | 0.865 | 0.493 | 0.622 | 0.187 | 0.647 |
| Kinase | 0.470 | 0.514 | 0.544 | 0.743 | 0.683 | 0.699 |
| Transf | 0.612 | 0.797 | 0.653 | 0.667 | 0.425 | 0.622 |
| Hydrol | 0.673 | 0.884 | 0.626 | 0.671 | 0.264 | 0.639 |
| EnzReg | 0.170 | 0.414 | 0.219 | 0.936 | 0.682 | 0.802 |

In the figure 3 can contrast the performance of computa­ tional consumption of SMOTE over another problems; thus this computational problem was execute on a intel i7 4th generation.

4. Conclusions

This paper allowed a comparison of two methods of balancing (SMOTE and Subsampling) using SVM with Gaussian kernel, opposite Weighted LSVM. Subsampling is known that the tool is not a reliable balance when protein sequences, by contrast with the results obtained are compa­ rable SMOTE of WSVM, this allows evaluating potential improvements.

As future work, the balancing methods can be tested over more complex scenarios like semi-supervised or multi­ label learning. Also, another strateggies such as Lagrangian SVM can be tested as well.

References

1. FEDECACAO. (2018). Retrieved August 9, 2018, from http://www.fedecacao.com.co/portal/index.php/es/2015-04-23-20-00-33/551-en-2017colombia-alcanzo-nuevo-record-en-produccion-de-cacao
2. GobernaGobernación de Santander. (2010). Plan de Desarrollo Departamental Santander Nos Une; 2016 - 2019. *PND Todos Por Un Nuevo País*, *1*, 419. <https://doi.org/10.1017/CBO9781107415324.004>
3. G. O. Consortium et al., "The gene ontology (go) database and informatics resource," Nucleic acids research, vol. 32, no. suppll,pp. D258-D261,2004.
4. J. A. Jaramillo-Garz6n, A. Perera-L1una,and G. Castellanos­ Dominguez, "Predictability of protein subcellular locations by pattern recognition techniques," in Proceedings of the

[Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/ summary?doi=IO.1.1.102.5233

[5] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of

imbalanced data: A review," International lournal of Pattern Recognition and Artificial Intelligence, vol. 23, no. 04, pp. 687-719,2009.

1. K. Veropoulos, C. Campbell, N. Cristianini, and Others, "Controlling the sensitivity of support vector machines," Proceedings of the international joint

conference on artificial intelligence, pp. 55-60, 1999.

[Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/ summary?doi=1 0.1.1.42. 7895$\ deli miter "026E30F$nhttp: IIseis. bri s.ac. uk/�enicgc/pubs/1999/ijcai \_ss. pdf

1. S. Garcia-L6pez,J. A. Jaramillo-Garz6n,and G. Castellanos­ Dominguez, "Optimization of cost sensitive models to im­

prove prediction of molecular functions," in Biomedical En­ gineering Systems and Technologies. Springer, 2014, pp. 207-222.

1. N. Y. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," lournal of artificial intelligence research, pp. 321-357,2002.
2. J. A. Jaramillo-Garz6n, J. J. Gallardo-Chac6n, C. G. Castellanos-Dominguez, and A. Perera-L1una, "Predictabil­ ity of gene ontology slim-terms from primary structure infor­ mation in embryophyta plant proteins," BMC bioinformatics, vol. 14,no. 1,p. 68,2013.

32nd Annual International Conference of the IEEE EMBS.

IEEE, August-September 2010, pp. 5512-5515.

1. C. Cortes and Y. Vapnik, "Support-vector networks," Ma­ chine learning, vol. 20,no. 3, pp. 273-297, 1995.
2. R. Akbani, S. Kwek, and N. Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets," Machine Learning: ECML 2004, vol. 3201, pp. 39-50, 2004.
3. C. Ceballes-Serrano, S. Garcia-L6pez, J. Jaramillo-Garz6n, y. Castellanos-Dominguez et aI., "A strategy for classifying imbalanced data sets based on particle swarm optimization,"

in Image, Signal Processing, and Artificial Vision (STSIVA),

2012 XVII Symposium of IEEE,2012,pp. 218-222.

1. Y. Tang, Y. Q. Zhang, and N. Y. Chawla, "SVMs modeling for highly imbalanced classification," IEEE Transactions on

4

Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 39,

no.l,pp. 281-288,2009.

1. C.-c. Chang and c.-J. Lin, "Libsvm: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2,no. 3,p. 27,2011.
2. E. Jain, A. Bairoch, S. Duvaud, I. Phan, N. Redaschi, B. E. Suzek,M. 1. Martin,P. McGarvey,and E. Gasteiger, "Infras­ tructure for the life sciences: design and implementation of the uniprot website," BMC bioinjormatics, vol. lO, no. 1, p. 136,2009.
3. J. Jaramillo-Garz6n, J. Gallardo-Chac6n, C. Castellanos­ Dominguez, and A. Perera-Lluna, "Predictability of gene ontology slim-terms from primary structure information in Embryophyta plant proteins," BMC Bioinjormatics, vol. 14, no. 1, p. 68, 2013. [Online]. Available: http: Ilwww.biomedcentral.comI1471-2105/14/68

5