Selección de características mediante la combinación de métodos para evaluar la precisión de clasificación en un conjunto de datos de implantes dentales

Nancy B. Ganz[1,*], Facundo A. Domínguez[2], Alicia E. Ares[1] y Horacio D. Kuna[2]
[1] Instituto de Materiales de Misiones, IMAM (CONICET-UNaM). Félix de Azara 1552, N3300LQH, Posadas, Misiones, Argentina.
[2] Departamento de Computación. Facultad de Ciencias Exactas, Químicas y Naturales. Universidad Nacional de Misiones. Félix de Azara 1552, N3300LQH, Posadas, Misiones, Argentina.
*E-mail: nancy.bea.ganz@gamil.com, facundokpo04@gmail.com, a.e.ares@gmail.com, hdkuna@gmail.com

RESUMEN

La selección de características es una técnica de preprocesamiento que permite encontrar un conjunto reducido de características, el cual concentra la información más sustancial del conjunto de datos. En este trabajo, se propone un procedimiento para la selección de las características más relevantes de un conjunto de datos de implantes dentales, de la Provincia de Misiones, Argentina. Se basa en la combinación de los métodos Information Gain, Gain Ratio, Random Forest importance, Relief y Chi Squared con el fin de predecir la clase minoritaria (Fracaso). El rendimiento del procedimiento propuesto se evaluó no sólo mediante la precisión de clasificación, en cuanto a las medidas de rendimiento tnr y bac de los clasificadores SVM rbf y Naive Bayes con validación cruzada, sino que también en base a la cantidad de características seleccionadas. Se observó que el procedimiento propuesto seleccionó la cantidad de características más adecuado para el estudio de caso y mejoró la precisión en la clasificación para la clase minoritaria.

Palabras Clave: *Ganancia de Información, Métodos, Selección de Características, Fracaso, Implantes Dentales.*

CONTEXTO

Esta línea de investigación se lleva a cabo dentro del Programa de Materiales y Fisicoquímica (PROMyF) en el Laboratorio de Ciencia de los Materiales del Instituto de Materiales de Misiones (IMAM), de la Facultad en Ciencias Exactas, Químicas y Naturales (FCEQyN), de la Universidad Nacional de Misiones (UNaM), en el marco de un plan de tesis doctoral, bajo el nombre de *"Aplicación de la Minería de Datos para la Selección de Biomateriales"*. Está financiado por el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) a través de una "Beca Interna Doctoral" otorgada por Resolución D Nº 4869. Además, cuenta con el auspicio del Colegio de Odontólogos de la Provincia de Misiones.

1. INTRODUCCIÓN

Este trabajo presenta un enfoque para la selección de características basado en la combinación de cinco métodos, para encontrar el subconjunto de características más relevante, evaluando la calidad de precisión en la clasificación y el número de características seleccionadas por cada método. La experimentación consistió en obtener los valores de importancia de cada característica en función de la integración de los métodos de selección de características: Information Gain[1], Gain Ratio[2], Random Forest importance[3], Relief[4] y Chi Squared[5]. Los pasos realizados fueron básicamente tres: generación del subconjunto de características, obtención de las medidas de rendimiento y apreciación de esas medidas para contrastar con el procedimiento propuesto. Para el propósito de este trabajo, se utilizó un conjunto de datos de historias clínicas de pacientes que se han sometido a procesos quirúrgicos de colocación de implantes dentales en la Provincia de Misiones, Argentina. Este conjunto de datos se encuentra representado a través de 4 dimensiones: datos del paciente

(antecedentes y condiciones médicas de los pacientes a la hora de la intervención), datos del implante (características del implante utilizado por el especialista implantólogo), datos de la fase quirúrgica (procesamiento de intervención quirúrgica y mejoramiento del lecho óseo del paciente) y datos del seguimiento postoperatorio (resultado del proceso de colocación del implante, es decir si el proceso de oseointegración implante/tejido-óseo tuvo éxito o fracasó). Para lograr este conjunto de datos se siguió un proceso, empleando la metodología CRISP-DM. Donde, en la primera fase correspondiente a la comprensión del problema, se evaluó la necesidad de la selección de características en conjuntos de datos desbalanceados. En la segunda fase se exploraron y estudiaron los datos, dimensiones, peculiaridades y verificación de calidad de los mismos, junto a los expertos. En la tercera fase, se procedió a la preparación y limpieza de los datos. Luego, se planteó un procedimiento para la selección de características, se probó y evaluó. Los trabajos[6]–[12] utilizan estos métodos de selección de características para reducir la dimensionalidad de los datos. Así mismo, utilizan comúnmente como clasificadores a Naive Bayes[13] y SVM[14] con validación cruzada. Y a la hora de medir la precisión de estos clasificadores, utilizan medidas de rendimiento como: matriz de confusión[15], accuracy, precisión y curva roc[16]. Es por esto que se ha tenido en cuenta estas particularidades para el diseño y validación del procedimiento propuesto, además de propiciar la combinación de los cinco métodos de selección de características propuestos para ofrecer la máxima precisión posible y no sesgar la decisión sobre los resultados de uno solo.

Debido a que la experimentación del procedimiento se realizó sobre un conjunto de datos desbalanceado, se buscó validar específicamente las métricas de rendimiento tnr y bac, debido a que son las medidas de referencia al tratarse de un conjunto de datos desbalanceado[17], [18].

Conjuntamente, se incorporó el análisis del método de selección de características que se extiende del algoritmo Random Forest, el cual se denomina Random Forest importance.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Esta línea de investigación propone estudiar, evaluar y aplicar distintas técnicas de ciencia de datos, para detectar los factores de fracaso de una base de datos de implantes dentales de la Provincia de Misiones. A través de la presente línea de investigación, se logró confeccionar el conjunto de datos necesario para proceder al diseño de un procedimiento, que a través de una metodología hibrida, permita identificar los factores de fracaso. Así mismo, se reconoce como objetivo de interés, el estudio de los tipos de tratamientos de superficie de los implantes utilizados en la zona, para incorporar como característica a evaluar.

3. RESULTADOS OBTENIDOS

Entre los resultados obtenidos, destacamos la creación de un procedimiento que consistió en:

Paso 1. Leer el conjunto de datos.

Paso 2. Seleccionar la característica objetivo para la predicción.

Paso 3. Obtener los subconjuntos de características de los métodos:

Information Gain (IG):

$IG(X) = H(Y) - H(Y|X) = H(X) - H(X|Y)$

Gain Ratio (GR):

$GR = \frac{IG}{H(X)}$

Random Forest importance (RFI):

$$RFI(x_m) = \frac{1}{N_T}\sum_{T}\sum_{t\epsilon T:v(S_t)=X_m} \mathrm{p(t)}\Delta i(\mathrm{s_t}, \mathrm{t})$$

Relief (R):

$$R_f = P\left(\frac{\text{different value of f}}{\text{class different}}\right) - P\left(\frac{\text{different value of f}}{\text{same class}}\right)$$

Chi Squared (ChiS):

$X^2 = \sum_{i=1}^{r}\ \sum_{j=1}^{c}\ \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

Paso 4. Confeccionar una matriz que concentre el valor de importancia obtenido por los distintos métodos para cada característica.

Paso 5. Normalizar los valores, debido a que los métodos empleados se desempeñan con rangos diferentes, este paso es necesario y fundamental para lograr un valor medio para cada característica. Se utilizó la función *normalize*, esta permite normalizar valores en base al método mínimo-máximo. La normalización mínimo-máximo regulariza las características en un rango[19]. Dado min_A y max_A valores mínimo y máximo de una característica A.

La normalización mínimo-máximo mapea un valor v_i de A para v_i' en el rango $[new\_min_A, new\_max_A]$ mediante $v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$.

Se utilizó este criterio de normalización debido a que este permite preservar todas las relaciones de los valores de los datos originales, es decir no introduce ningún sesgo potencial en los datos. Además, se encuentra demostrado que tiene mejor rendimiento en la clasificación[20], [21]. El rango empleado fue [0,1].

Paso 6. Transponer la matriz.

Paso 7. Obtener la mediana de cada característica en función de los valores arrojados por los distintos métodos. Se empleó la mediana como medida de tendencia central debido a que los valores de importancia arrojados por los distintos métodos no seguían una distribución normal. En el caso de que estos valores sigan una distribución normal aplicar la media[19].

Paso 8. Ordenar la matriz en forma decreciente.

Paso 9. Obtener un umbral, a través de validación cruzada de 10 iteraciones. De esta manera se logró hallar el umbral más óptimo para las características cuyos valores de importancia resultaron del paso 7.

Paso 10. Seleccionar las características que cumplan con la condición de ser igual o mayor al umbral obtenido en el paso 9.

Paso 11. Obtener un umbral óptimo para cada uno de los 5 métodos, esto se realizó a través de una calibración del método sobre el conjunto de datos con un clasificador Naive Bayes y una validación cruzada de 10 veces. Es necesario aclarar, que el umbral conjuntamente fue ajustado en función del conocimiento que se tiene sobre los datos.

Paso 12. Seleccionar las características para cada uno de los 5 métodos que cumplan con la condición de ser igual o mayor al umbral obtenido en el paso 11.

Paso 13. Aplicar los clasificadores SVM $f(X) = W^T\emptyset(X)$ con núcleo rbf $K(x_i, x_j) = \emptyset\,(x_i).\emptyset(x_j)$ y Naive Bayes $f_i(X) = \prod_{j=1}^{N} P(x_j|c_i)\,P(c_i)$ con una validación cruzada de 100 iteraciones sobre los 5 conjuntos de características obtenidos en el paso 12 y a la matriz normalizada lograda en el paso 10.

Paso 14. Obtener las medidas de rendimiento: tpr, fpr, tnr, fnr, bac, auc y acc.

Paso 15. Validar las medidas tnr y bac.

Cabe aclarar, que tanto a los métodos IG, GR, RFI, R y ChiS, así como a los clasificadores, se los consideró a cada uno con el mismo peso. Debido a que la finalidad fue optimizar la cantidad y conocer que características seleccionaba cada método, para posteriormente corroborar su rendimiento individual con los clasificadores.

Para la funcionalidad del procedimiento propuesto se utilizó un conjunto de datos de historias clínicas de implantes dentales, con 1.050 filas, 31 características y un atributo clase binario: EXITO (977 casos) y FRACASO (73 casos).

La implementación se efectuó sobre la herramienta RStudio versión 3 de GNU con licencia AGPL (Affero General Public License), cuyo software es gratuito y de código abierto[22]. Conjuntamente, se utilizó la librería MLR para disponer de los métodos de selección de características. Este es un paquete muy completo, el cual proporciona métodos supervisados como clasificación y regresión. Así mismo, provee métodos no supervisados como agrupación, junto con métodos de evaluación y optimización[23]. Al mismo tiempo, el paquete MLR utiliza los paquetes FSelector[24] y randomForest[25] para obtener los algoritmos de selección de características.

La Tabla 1 muestra la cantidad de características seleccionadas por los métodos utilizados y el procedimiento propuesto (PP), así como los resultados arrojados con los clasificadores.
En cuanto a los valores arrojados de la clasificación, se observó que comparando con los métodos IG, GR, RFI, R y ChiS, PP logró la mejor tasa de verdaderos negativos (tnr) con el clasificador Naive Bayes y una medida bac de 53%.

Tabla 1 – Resultados obtenidos de aplicar los métodos IG, GR, RFI, R, ChiS y el procedimiento propuesto con los clasificadores SVM rbf y Naive Bayes.

Métodos	Cant.	tpr	fpr	tnr	fnr	bac	auc	acc
IG-SVM	10	0.997	0.989	0.011	0.003	0.504	0.698	0.929
IG-NB		0.982	0.949	0.051	0.018	0.517	0.720	0.917
GR-SVM	11	0.999	1.000	0.000	0.001	0.500	0.563	0.930
GR-NB		0.992	0.937	0.063	0.008	0.527	0.726	0.927
RFI-SVM	22	0.998	0.938	0.062	0.002	0.530	0.680	0.933
RFI-NB		0.974	0.958	0.042	0.026	0.508	0.694	0.909
R-SVM	18	0.998	0.978	0.022	0.002	0.510	0.676	0.930
R-NB		0.981	0.916	0.084	0.019	0.533	0.697	0.919
ChiS-SVM	16	0.998	0.999	0.001	0.002	0.500	0.678	0.929
Chis-NB		0.975	0.918	0.083	0.025	0.529	0.734	0.913
PP-SVM	17	0.996	0.974	0.026	0.004	0.511	0.710	0.929
PP-NB		0.975	0.913	**0.087**	0.025	**0.531**	0.732	0.913

Basándonos en los resultados, podemos afirmar que los algoritmos de selección de características utilizados, si se los emplea de forma individual, pueden ocasionar una incorrecta selección de subconjuntos de características. Una adecuada combinación de estos métodos, puede producir subconjuntos de características más efectivas para la clasificación, como ocurre en este caso.
También hemos contemplado que, los métodos de selección de características IG y GR lograron reducir en ocasiones más la cantidad de características que los métodos RFI, R, ChiS y PP, pero su precisión de clasificación con los clasificadores SVM rbf y NB no fueron tan buenos. Además, PP seleccionó la cantidad más adecuada de características, siendo éstas las más efectivas en la clasificación.
Se observó, que el clasificador NB fue superior respecto a SVM, debido a que logró mejores rendimientos, y sobre todo al tratarse de un conjunto de datos desbalanceado. Conjuntamente, se valoró que SVM requiere de mayor calibración y necesita conocer a priori los datos, para mejorar su rendimiento y obtener resultados comparables.
Por lo tanto, podemos afirmar que PP es útil cuando se requiere definir aquellas características que se encuentran en un border line, además combina varios métodos y no sesga la decisión sobre los resultados de una sola técnica. También, en base a la opinión de los especialistas, se pudo validar que los factores detectados son los que mayor influencia ejercen sobre el proceso de oseointegración.

4. FORMACIÓN DE RECURSOS HUMANOS

Este proyecto es parte de las líneas de investigación del "Programa de Materiales y Fisicoquímica" de la FCEQyN – UNaM, con cuatro integrantes: un investigador independiente del CONICET, un docente categoría I perteneciente al Depto. de Informática de la FCEQyN – UNaM, un

doctorando y un tesista de grado de la carreara de Licenciatura en Sistemas de Información.

5. BIBLIOGRAFÍA

[1] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[2] J. R. Quinlan, "Induction of Decision Trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[3] L. Breiman, "Random Forest," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[4] K. Kira and L. A. Rendell, "A Practical Approach to Feature Selection," *Mach. Learn. Proc.*, pp. 249–256, 1992.

[5] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 50, no. 302, pp. 157–175, 1900.

[6] A. Chaudhary, S. Kolhe, and Rajkamal, "Performance Evaluation of feature selection methods for Mobile devices," *J. Eng. Res. Appl.*, vol. 3, no. 6, pp. 587–594, 2013.

[7] Z. Karimi, M. Mansour, R. Kashani, and A. Harounabadi, "Feature Ranking in Intrusion Detection Dataset using Combination of Filtering Methods," *Int. J. Comput. Appl.*, vol. 78, no. 4, pp. 21–27, 2013.

[8] L. Gao, M. Ye, X. Lu, and D. Huang, "Hybrid Method Based on Information Gain and Support Vector Machine for Gene Selection in Cancer Classification," *Genomics, Proteomics Bioinforma.*, vol. 15, no. 6, pp. 389–395, 2017.

[9] T. Z. Phyu and N. N. Oo, "Performance Comparison of Feature Selection Methods," *MATEC Web Conf.*, vol. 42, p. 06002, 2016.

[10] H. Dag, K. E. Sayin, I. Yenidogan, S. Albayrak, and C. Acar, "Comparison of Feature Selection Algorithms for Medical Data," *Int. Symp. Innov. Intell. Syst. Appl.*, pp. 1–5, 2012.

[11] M. Peker, A. Arslan, B. Sen, F. V. Celebi, and A. But, "A novel hybrid method for determining the depth of anesthesia level: Combining ReliefF feature selection and random forest algorithm (ReliefF+RF)," *Int. Symp. Innov. Intell. Syst. Appl.*, 2015.

[12] W. Li, Y. Li, J. Chen, and C. Hou, "Product functional information based automatic patent classification: Method and experimental studies," *Inf. Syst.*, vol. 67, pp. 71–82, Jul. 2017.

[13] N. Friedman, D. Geiger, and M. Goldszmit, "Bayesian Network Classifiers," *Mach. Learn.*, vol. 29, pp. 131–163, 1997.

[14] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[15] R. Susmaga, "Confusion Matrix Visualization," in *Intelligent Information Processing and Web Mining*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 107–116, 2004.

[16] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[17] Q. Wei and R. L. Dunbrack, "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics," *PLoS One*, vol. 8, no. 7, 2013.

[18] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[19] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[20] W. Li and Z. Liu, "A method of SVM with normalization in intrusion detection," *Procedia Environ. Sci.*, vol. 11, no. PART A, pp. 256–262, 2011.

[21] S. Jain, S. Shukla, and R. Wadhvani, "Dynamic selection of normalization techniques using data complexity measures," *Expert Syst. Appl.*, vol. 106, pp. 252–262, 2018.

[22] "RStudio - RStudio." [Online]. Available: https://www.rstudio.com/products/rstudio/. [Accessed: 26-Feb-2019].

[23] B. Bischl *et al.*, "mlr: Machine Learning in R," *J. Mach. Learn. Res.*, vol. 17, pp. 1–5, 2016.

[24] P. Romanski and L. Kotthoff, "Package 'FSelector,'" *Repository CRAN*. 2018.

[25] L. Breiman and A. Cutler, "Package 'randomForest,'" *Repository CRAN*. 2018.