Simulación de Sistemas y Laboratorio: Entrega Final Proyecto

Jiménez, Sebastián. Londoño, Jaime.
sebasj14@gmail.com jaime951@gmail.com
Universidad de Antioquia

Resumen— Se abordará un problema de clasificación donde se evalúa la esperanza de vida de un paciente de cirugía de tórax debido a cáncer de pulmón. Se aplicarán los métodos de Discriminantes Gaussianas, K-vecinos más cercanos, Random Forest, Redes Neuronales Artificiales y Máquinas de Soporte Vectorial, con el fin de encontrar el más eficiente para este problema particular. Posteriormente, se implementarán técnicas de selección de características como Búsqueda Secuencial Ascendente (SFS), Análisis de Componentes (PCA) y Análisis Discriminante (LDA). En cada método se usará Cross-Validation como técnica de validación y se aplicaran técnicas de submuestreo y sobremuestreo inteligente para tratar los problemas de desbalance de los datos.

Índice de Términos— Clasificación, Cirugía de tórax, Discriminantes Gaussianas, Redes Neuronales Artificiales, SMV, PCA, SFS, LDA, Knn, Cross-Validation, submuestreo, sobremuestreo.

I. INTRODUCCIÓN

En este artículo se presentan modelos de predicción para clasificar la esperanza de vida de pacientes de cáncer de pulmón luego de ser sometidos a cirugía. Esto es, de acuerdo a una serie de características del paciente, diagnosticar si sobrevive más de un año luego de la cirugía. El objetivo es definir la mejor técnica entre las siguientes para abordar este problema en particular: Discriminantes Gaussianas, K-vecinos más cercanos, Random Forest, Redes Neuronales Artificiales y Máquinas de Soporte Vectorial.

El sistema que salga definido como el más eficiente, puede ser bastante útil en el área de la salud, específicamente en aquellos departamentos encargados de evaluar la condición de personas que han sido sometidas a cirugía debido a cáncer de pulmón. Área en la que en los datos que registra suele presentarse desproporción en la distribución de clases, lo que se conoce como problema de desbalanceo de datos.

Por otro lado, se aplican técnicas de selección de características, buscando reducir la dimensión de la base de datos. Esto con el objetivo de reducir el costo computacional que puede demandar el procesamiento de una base de datos con una cantidad de variables muy alta, ya que se eliminan las variables que menos aportan al problema.

Para este modelamiento, se utiliza un conjunto de datos brindado por el repositorio para aprendizaje de máquinas de la Universidad de Irvine. (UCI Machine Learning Repository) llamado Thoracic Surgery Data [1].

II. TRABAJOS RELACIONADOS

Existen, en la literatura científica, pocos aportes conocidos que aborden problemas con la misma base de datos.

En [2] se presenta la técnica de Máquina de Vectores de Soporte (SVM) para solucionar el problema del desbalance en la base de datos. Esta solución propuesta combina los beneficios de usar clasificadores de ensamble para datos desiguales, junto con Máquinas de Vectores de Soporte Sensibles al Costo (CSVM).
Para la evaluación del sistema, se utiliza la técnica Stratified Cross-Validation, que consiste en conseguir que en cada subconjunto en el que se divide el total de muestras, haya igual número de ejemplares de cada clase representada.
Entre los resultados obtenidos en este trabajo, destacan el uso de la técnica UnderBagging que obtuvo un 68.57% en True Positive, y un 100% en True Negative, usando la técnica JRip.

En [3] se presentan diferentes técnicas de aprendizaje con las que se aborda el problema de predicción de la esperanza de vida de pacientes de cirugía por cáncer de pulmón. Entre ellas están Naive Bayesian Classifier, PART (Partial Decision Trees), J48 Decision Tree, OneR Classifier, Random Forest y Decision Stump.
Como técnica de evaluación de los sistemas, se utiliza K-fold Cross-Validation. En el artículo se concluye que la mejor técnica para resolver este problema es el Random Forest, que devuelve un 95.65% de precisión.

En [4] se trabaja con el mismo problema y la misma base de datos, donde se reconoce también el problema de desbalance de los datos. Abordan el problema implementando la técnica Ensemble SVM (ESVM).
Además, se usa Cross-Validation de 10 folds con 10 repeticiones como técnica de evaluación del sistema. Y por último comparan los resultados obtenidos con los resultados

de usar SVM entrenado usando SMO, y con los resultados de Cost-Sensitive SVM entrenado con SMO modificado. De esto, se deja muestra que si bien la precisión fue más baja (60.37%), el valor de GMean y el porcentaje de las muestras de la clase positiva detectadas correctamente fueron mayores que en los otros dos métodos (47.33% de Gmean y 33.79% de TruePositive).

En [5], si bien no se trabaja con la misma base de datos, sí se trabaja un problema similar, y su enfoque está en una base de datos para detectar cáncer de próstata. Por otro lado, en [6] se enfocan en diferentes técnicas para afrontar el desbalanceo en diferentes conjuntos de datos que plantean un acercamiento diferente a este tipo de problemas.

III. EXPERIMENTOS

Como se estableció previamente, en este artículo se trabaja el problema de clasificación que plantea la base de datos Thoracic Surgery [1]. A continuación se enumeran las variables de entrada que posee el sistema:

1. DGN: Diagnóstico— Combinación específica del código ICD-10 para identificar el tipo de tumor. (DGN3,DGN2,DGN4,DGN6,DGN5,DGN8,DGN1)
2. PRE4: Capacidad Vital Forzada - FVC (numérico)
3. PRE5: Volumen que ha sido exhalado al final del primer segundo de espiración forzada - FEV1 (numérico)
4. PRE6: Calidad de vida del paciente - escala Zubrod (PRZ2,PRZ1,PRZ0)
5. PRE7: Dolor antes de la cirugía (T,F)
6. PRE8: Hemoptisis antes de la cirugía (T,F)
7. PRE9: Disnea antes de la cirugía (T,F)
8. PRE10: Tos antes de la cirugía (T,F)
9. PRE11: Estado débil antes de la cirugía (T,F)
10. PRE14: T en el sistema de estadificación TNM - tamaño del tumor primario, siendo OC11 el más pequeño y OC14 el más grande (OC11,OC14,OC12,OC13)
11. PRE17: El paciente presenta Diabetes Mellitus Tipo 2 (Diabetes del adulto) (T,F)
12. PRE19: El paciente ha presentado Infarto Agudo de Miocardio en los últimos 6 meses (T,F)
13. PRE25: El paciente presenta Enfermedad Arterial Periférica (EAP). (T,F)
14. PRE30: El paciente fuma (T,F)
15. PRE32: El paciente presenta Asma (T,F)
16. AGE: Edad del paciente al momento de la cirugía (numérico)

Mientras que la variable de salida o variable a predecir es la siguiente:
1. Risk1Y: Periodo de supervivencia de un año - Valor (T)rue si murió (T,F).

Se cuenta con 470 muestras, donde 400 pertenecen a la clase etiquetada “negativa” y 70 de la “positiva”, lo cual deja en evidencia el problema de desbalance tan grande que posee la base de datos.

Este problema fue tratado aplicando técnicas de sobremuestreo inteligente sobre la clase minoritaria y submuestreo inteligente sobre la clase mayoritaria, logrando así a un mismo tiempo crear muestras artificiales de una clase y eliminar los datos ruidosos o atípicos de la otra.

Luego de aplicar las técnicas de submuestreo en la clase mayoritaria, se eliminó un 16% de los datos, lo que dejó 336 muestras de tal clase, cuatro veces más que la clase minoritaria luego del sobremuestreo, que fue del 20%. Es decir, se crearon 14 muestras artificiales, para un total de 84.

Se utilizaron varios experimentos para la predicción de la esperanza de vida de los pacientes. En primer lugar está la técnica de Funciones Discriminantes Gaussianas como modelo paramétrico. Segundo, está la técnica de los k-vecinos más cercanos (KNN). Luego, el modelo Random Forest, Redes Neuronales Artificales y por último el uso de Máquinas de Vectores de Soporte.

En todos estos casos se usa Cross-Validation como método de validación, mientras que se tomarán como criterio de validación el error de clasificación, la especificidad, la sensibilidad y porcentaje de falsos negativos, ya que nos interesa más que el sistema tenga menos porcentaje de pacientes que se diga que mueren dentro de un año luego de la cirugía, pero el sistema los clasifica como que sobreviven, pues existe la posibilidad de que un paciente esté tranquilo debido a que no morirá en el primer año, y que sí muera en ese periodo.

IV. RESULTADOS

Como se estableció previamente, se tomarán como criterios para elegir el mejor modelo la sensibilidad, la especificidad, la eficiencia, el intervalo de confianza de la eficiencia y la media geométrica. Se registra en la tabla 1 el mejor resultado de cada modelo usado (Funciones Discriminantes Gaussianas, K-Vecinos, Redes Neuronales Artificiales, Random Forest y Máquinas de Soporte Vectorial.

TABLA I
MEJOR RESULTADO DE CADA MODELO

	F. D. Gaussianas	K-Vecinos	Redes Neuronales	Random Forest	Máquinas de Vectores
Parámetro	N/A	3 Vecinos	13 Percept.	10 Árboles.	Kernel Lineal B. Const = 1
Sensibilidad	0.2	0.1375	0.188	0.1172	0.6026
Especificidad	0.8	0.9034	0.947	0.9519	0.6639
Eficiencia	0.69362	0.7893	0.834	0.8255	0.6553
IC Eficiencia	+- 0.31434	+- 0.0406	+- 0.0457	+- 0.0591	+- 0.0861
Med. Geomet.	0.4	0.3524	0.422	0.334	0.6325
Tiempo de Entrenam.	0.030311 seg	0.080425	0.768984 seg	0.221991 seg	1.172025 seg
Tiempo de Validación	0.000748 seg		0.014385 seg	0.040302 seg	0.002478 seg

Teniendo en cuenta la tabla de resultados, se optó por dejar como mejor criterio de selección la sensibilidad, ya que nos interesa más que el sistema tenga un bajo porcentaje de personas que se diga que mueren dentro de un año luego de la cirugía, pero el sistema los clasifica como que sobreviven, pues existe la posibilidad de que un paciente esté tranquilo debido a que se dijo que no morirá en el primer año, y que sí muera en ese periodo.

El mejor resultado se encontró con el modelo de Máquinas de Vectores de Soporte con Kernel Lineal y Box Costraint = 1, dado que se encontraron buenos valores tanto en eficiencia como en sensibilidad (ambos por encima de 60%). Sin embargo, el tiempo de ejecución (en entrenamiento y validación) es mucho mayor que el de los demás modelos.

V. REDUCCIÓN DE DIMENSIÓN

Cuando un conjunto de datos posee una dimensión muy alta se pueden presentar complicaciones en cuanto al costo computacional que demanda procesar tal cantidad de variables. Por otro lado, en muchos casos, no todas las variables registradas en un conjunto de datos son relevantes para el problema en cuestión. Esto es, algunas variables no aportan información valiosa para explicar el problema de interés. [9]

Debido a esto, se hace uso de técnicas para disminuir la cantidad de características de un conjunto de datos. Este proceso es llamado Reducción de Dimensión.

En este artículo, se aplicarán tres técnicas para dicha área sobre la base de datos Thoracic Surgery: Búsqueda Secuencial hacia adelante (SFS), Análisis de Componentes Principales (PCA) y Análisis Discriminante Lineal (LDA).

A. Búsqueda Secuencial hacia Adelante

Este es un algoritmo que incluye el mejor atributo determinado por una función objetivo. Luego, de los atributos restantes, se elige el mejor, y así sucesivamente hasta que no se obtenga un mejoramiento al agregar un atributo. Normalmente, esta técnica es usada con la evaluación tipo wrapper, donde la eficiencia es el criterio para determinar las mejoras [10]

Para la evaluación tipo wrapper se usan como criterios los tres modelos que obtuvieron los mejores resultados en la sección Resultados de este artículo (ver Tabla 1). Luego de haber aplicado esta técnica, se obtuvo que las características más relevantes son:
Característica 5: Dolor antes de la cirugía.
Característica 6: Hemoptisis antes de la cirugía.
Característica 11: El paciente presenta Diabetes Mellitus T2.
Característica 15: El paciente presenta asma.

Para evaluar estas cuatro características, se encuentran nuevamente todos los criterios escogidos en la sección anterior (Sensibilidad, Especificidad, etc – Ver Tabla 1). Esto es, se entrenan y validan los mejores modelos con los mejores parámetros y solo con las características seleccionadas. Los resultados se consignan en la Tabla 2.

TABLA II
RESULTADOS MEJORES MODELOS LUEGO DE SFS

	F. D. Gaussianas	Redes Neuronales	Máquinas de Soporte Vectorial
Parámetro	N/A	13 Perceptrones	Kernel Lineal B. Const = 1
Sensibilidad	0.2	0.0541	0.8724
Especificidad	0.8	0.9828	0.0964
Eficiencia	0.7191	0.8447	0.2170
IC Eficiencia	+-0.2946	+-0.438	+-0.1980
Media Geométrica	0.4	0.2306	0.2899

A partir de la tabla, se puede concluir que la sensibilidad de las Máquinas de Soporte mejoró considerablemente, pero la eficiencia se vio afectada negativamente, lo que conllevó a que la Media Geométrica también fuera relativamente baja.

En los otros dos modelos no se observa un cambio significativo.

Cabe aclarar que este resultado se obtuvo al tener en cuenta 3 componentes.

B. Análisis de Componentes Principales

Es una técnica de extracción de características que sirve para reducir la dimensionalidad de un conjunto de datos con un largo número de variables correlacionadas, mienta mantiene en el mayor grado posible la variación presente en los datos. Dicho objetivo es logrado transformando a un nuevo conjunto de variables, los componentes principales, los cuales son ordenados de manera que el primer componente conserva la mayor parte de la variación presente en las variables originales y las demás componentes no están correlacionados. [8]

Los resultados al aplicar esta técnica están consignados en la Tabla 3.

TABLA III
RESULTADOS MEJORES MODELOS CON PCA

	F. D. Gaussianas	Redes Neuronales	Máquinas de Soporte Vectorial
Parámetro	N/A	13 Perceptrones	Kernel Lineal B. Const = 1
Sensibilidad	0.6667	0	0.75
Especificidad	0.3158	1	0.5641
Eficiencia	0.3830	0.8723	0.5957
IC Eficiencia	+-1.3879	+-0-1957	+-0.0456
Media Geométrica	0.4588	0	0.6504

Esta técnica propone una mejora en Máquinas de Soporte con respecto a los resultados sin aplicar técnicas de selección/extracción. Constituyéndose como mejor técnica que

SFS, debido a la mejor relación entre Sensibilidad y Eficiencia. Lo que se ve respaldado por el mejor valor de Media Geométrica.

Los resultados registrados corresponden al obtener 3 componentes principales.

C. Análisis Discriminante Lineal

Es una técnica de aprendizaje supervisado para clasificar datos. Su funcionamiento consiste en proyectar los datos en un espacio de menor o incluso igual dimensión que los datos de entrada, donde se busca que las clases en la proyección estén lo más separadas posibles, evitando su solapamiento. Esto último se diferencia de PCA, donde esto no se garantiza.

Esta técnica propone normalizar la distancia entre las medias de dispersión, de forma que se consigue una clasificación más precisa, ya que se pueden obtener variables cuyas medias están alejadas y no están solapadas sus distribuciones.

En la tabla 4 están registrados los resultados arrojados luego de aplicar esta técnica.

TABLA IV
RESULTADOS MEJORES MODELOS CON LDA

	F. D. Gaussianas	Redes Neuronales	Máquinas de Soporte Vectorial
Parámetro	N/A	13 Perceptrones	Kernel Lineal B. Const = 1
Sensibilidad	0.7143	0.3125	0.7692
Especificidad	0.6770	0.9359	0.4815
Eficiencia	0.6809	0.8298	0.5213
IC Eficiencia	+-0.0458	+-0.5667	+-1.4326
Media Geométrica	0.6944	0.5408	0.6086

Los resultados registrados corresponden al obtener 3 componentes principales. Se observa una mejora considerable en el modelo de Máquinas de Soporte Vectorial con respecto a la técnica SFS y PCA. Si bien disminuye la Sensibilidad, tanto la Eficiencia, como Especificidad y Media Geométrica aumentan en grandes valores.

En cuanto a Funciones Discriminantes Gaussianas, se consiguieron los mejores resultados en todos los criterios de todas las técnicas.

Por otro lado, el mejor resultado al aplicar esta técnica se encuentra en las Funciones Discriminantes Gaussianas, con una Media Geométrica de casi 70%, demostrando su eficiencia en la predicción en ambas clases.

VI. DISCUSIÓN

En general, en los modelos evaluados se obtuvo una eficiencia menor que en los casos de trabajos relacionados donde se trabaja con la misma base de datos. Como se decidió tomar la sensibilidad como el mejor criterio de validación, con un peso mayor incluso que el del error de clasificación, se le da más importancia a este resultado que a la eficiencia en sí. Esto es debido a que es más importante que el sistema evalúe mejor los pacientes que sobreviven en el período de un año después de la cirugía.

A pesar de que la eficiencia más alta fue obtenida por la técnica de Redes Neuronales (casi 85% con 13 perceptrones), no se escoge como el mejor modelo, ya que al verificar los valores de Sensibilidad, la técnica que mejor desempeño tuvo fue Funciones Discriminantes Gaussanas (Sensibilidad 71% - Eficiencia 68%, luego de aplicar la técnica LDA).

Por otra parte, la técnica que mejor sensibilidad consiguió fue Máquinas de Vectores de Soporte (Sensibilidad 87%, Eficiencia 22% con Kernel lineal y Box Constraint = 1). Este criterio es bastante importante en el campo de la medicina.

VII. CONCLUSIONES

La selección e implementación de la técnica más adecuada para modelar los datos es tan importante como el tratamiento del desbalance en las bases de datos que cuenten con este problema para obtener un bajo error de clasificación en todas las clases.

Como consecuencia de lo anterior, se evidencia que no siempre una alta eficiencia es la mejor medida de desempeño de un sistema, ya que esta no toma en cuenta el ajuste del modelo a las diferentes clases y puede ocultar los problemas de desbalance presentes en un conjunto de datos.

En aplicaciones médicas es importante conocer que tipos de error son más nocivos, dado que estos son los que se deben minimizar.

Se encontró que aplicando técnicas de extracción/selección de características, podría mejorarse el entendimiento de un problema dado que se sabe qué variables determinan el comportamiento de las salidas.

VIII. REFERENCIAS

[1] K. Bache y M. Lichman, «UCI Learning Machine Repository,» University of California, School of Information and Conputer Sciences, [En línea] Disponible en: https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data

[2] Maciej Zięba, et al. “Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients”. Applied Soft Computing, Volume 14, Part A, January 2014, Pages 99-108, ISSN 1568-4946, Disponible en: http://dx.doi.org/10.1016/j.asoc.2013.07.016

[3] Sindhu. V, et al. “Thoracic Surgery Analysis Using Data Mining Techniques”. Int.J.Computer Technology & Applications,Vol 5 (2),578-586. ISSN:2229-6093, Disponible en:

http://www.ijcta.com/documents/volumes/vol5issue2/ijcta2014050253.pdf

[4] Zieba, M.; Swiatek, J., "Ensemble SVM for imbalanced data and missing values in postoperative risk management," e-Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on , vol., no., pp.95,99, 9-12 Oct. 2013, Disponible en: 10.1109/HealthCom.2013.6720646

[5] Artan, Y.; Haider, M.A.; Langer, D.L.; van der Kwast, T.H.; Evans, A.J.; Yongyi Yang; Wernick, M.N.; Trachtenberg, J.; Yetik, I.S., "Prostate Cancer Localization With Multispectral MRI Using Cost-Sensitive Support Vector Machines and Conditional Random Fields," Image Processing, IEEE Transactions on , vol.19, no.9, pp.2444,2455, Sept. 2010 doi: 10.1109/TIP.2010.2048612

[6] Haibo He; Garcia, E.A., "Learning from Imbalanced Data," Knowledge and Data Engineering, IEEE Transactions on , vol.21, no.9, pp.1263,1284, Sept. 2009
doi: 10.1109/TKDE.2008.239

[7] The MathWorks, Inc., «MathWorks,» MathWorks Inc., 2014. [En línea]. Disponible en: http://www.mathworks.com/help/stats/treebagger.html

[8] I. Jolliffe, Principal Component Analysis, Encyclopedia of Statistics in Behavioral Science, 2005

[9] I. K. Fodor, «A survey of dimension reduction techniques,» 2002

[10] P. Bermejo, J. Gamez y J. Puerta, «Incremental Wrapper-based subset Selection with replacement: An advantageous alternative to sequential forward selection,» de *Computational Intelligence and Data Mining*, 2009