ESTUDIO COMPARATIVO DE TÉCNICAS SUPERVISADAS DE MACHINE LEARNING APLICADAS EN PROBLEMAS MÉDICOS

COMPARATIVE STUDY OF SUPERVISED TECHNIQUES OF MACHINE LEARNING APPLIED TO MEDICAL PROBLEMS

Basante Cielo[1], Ortega Carlos[1], Peluffo Diego[2] y Blanco Xiomara[3]

[1] Universidad de Nariño – Colombia, [2] Universidad Técnica del Norte – Ecuador y [3] Universidad de Salamanca– España.

Correo electrónico para correspondencia: {cielo.2103,carlosmaor}@udenar.edu.co,{dhpeluffo}@utn.edu.ec,{xiopepa}@usal.es

Resumen:

Actualmente la sobrecarga de información ha provocado que las capacidades humanas de análisis se queden cortas ante el inminente crecimiento de las capacidades tecnológicas para recolectar, comunicar y guardar grandes volúmenes de información. En los hospitales o entidades de salud se recolectan a diario millones de datos y registros de exámenes, pruebas diagnósticas, entre otras. La medicina ha utilizado la revolución tecnológica de muchas maneras para facilitar el manejo, procesamiento, visualización y análisis de datos. Una tarea en específico asociada al diagnóstico de enfermedades es la clasificación de patologías en donde la máquina entrenada en conjunto con el experto puede mejorar y facilitar el diagnostico. Actualmente existen muchas disciplinas que tiene como objetivo desarrollar técnicas que le permiten a las computadoras aprender, entre las que se encuentra machine learning, en la cual se han desarrollado diversos clasificadores supervisados que permiten la organización de nuevos objetos representados en características a categorías ya conocidas.

La clasificación patologías, como parte de la inteligencia artificial, no es una tarea fácil y más cuando existen una gran variedad de técnicas de clasificación, por tanto en este trabajo de investigación se realizará una comparación de técnicas de clasificación multi-clase recomendada por la literatura científica, tales como: SVM (Support Vector Machine), KNN (K-nearest neighbors algorithm), ANN (Artificial neural network), PC (Parzen´s Classifier), árboles de decisión (Random Forest) y Adaboost (Adaptative Boosting). Los experimentos se realizan sobre bases de datos de enfermedades cardiacas (Cleveland), cardiotocografía e hipotiroidismo.

Como resultado, se evidenció que las bases de datos Cleveland e hipotiroidismo fueron un reto para los clasificadores ANN, PC y Adaboost por su escasez de datos, provocando que la clasificación tenga un error considerable. Para bases de datos que tienen mejor equilibrado su número de muestras para cada clase como cardiotocografía se obtienen mejores resultados al ser entrenado con Adaboost.

Palabras clave: Machine learning, técnicas supervisadas, hipotiroidismo, cleveland, inteligencia artificial.

Abstract:

Currently, information overload has caused human analysis capabilities become insufficient in comparison with the imminent growth of technological capabilities to collect, communicate and store large volumes of information. In hospitals or health entities, millions of data, diagnostic tests, and other information are collected every day. The medicine has used the technological revolution in many ways to facilitate the handling, processing, visualization and analysis of data, a specific task associated with the diagnosis of diseases is the pathology classification, wherein the machine trained in conjunction with an expert can improve and facilitate the diagnosis. Currently there are many disciplines which aim is developed techniques that allow computers to learn, however one of the most explored is machine learning in which classifiers have been developed, they allow the assignment of new objects represented in features into beforehand known categories.

As part of artificial intelligence, the classification of pathologies is not a trivial task - there is a great variety of classification techniques. In this connection, this research work performs a comparison of multi-class classification techniques recommended by scientific literature, such as: SVM (Support Vector Machine), KNN (K-nearest neighbors algorithm), ANN (Artificial neural network), PC (Parzen´s Classifier), Random Forest and Adaboost (Adaptative Boosting). The experiments are performed on heart disease (Cleveland), cardiotocography and hypothyroidism databases.

As a result, it is demonstrated that Cleveland and hypothyroid databases are challenging for the ANN, PC and Adaboost classifiers because of their scarcity of data, causing that classification procedures reach greater error values. On the contrary, when dealing with databases having proper balance in their number of samples for each class (for instance, cardiotocography), better results are obtained when classification training is done with Adaboost.

Keywords: Machine learning, supervised techniques, hypothyroidism, cleveland artificial intelligence.

Introducción:

Con el enorme auge tecnológico y la revolución digital de la actualidad se ha hecho más común la necesidad de recolectar, procesar, almacenar y transmitir la información digitalizada, es así como día a día nos vemos en vueltos en la era digital o también llamada la era de los datos (Witten, Frank, Hall y Pal, 2016). Los volúmenes de datos vienen creciendo desde hace más de dos décadas y se evidencia en episodios simples de millones de empresas, hospitales, entidades, etc., quienes a diario capturan datos de usuarios, registran actividad o almacenan información en la red mundial.

La innovación y el crecimiento de la tecnología en el sector salud ayuda en la predicción y análisis de enfermedades ayudándole al experto a diagnosticarlas en una fase inicial, permitiendo la ejecución temprana de terapias o tratamientos médicos (Siegel, 2013). El diagnóstico es la clasificación del paciente en una patología en concreto, en otras palabras, se clasifica teniendo en cuenta una serie de valores o atributos característicos de una patología en particular que son medidos u observados por ejemplo talla, peso, temperatura, presión, etc.

La inteligencia artificial, aprendizaje automático y minería de datos son herramientas que facilitan el análisis de datos, en principio estos métodos no eran tan usados en el área de la medicina por razones culturales o filosóficas, las cuales asumen que una computadora nunca será tan capaz como un humano de acertar en el diagnóstico de una patología, esto aunado al hecho de que algunos médicos se sienten cuestionados, supervisados o aconsejados por una maquina o por un ingeniero (Reyes, Colín y Murata ,2014). A pesar de esto, lidiar con bases de datos bastante grandes y complejas con métodos tradicionales resulta complicado por lo que en los últimos años ha habido una gran acogida al desarrollo y adaptación de software para facilitar estas tareas.

Teniendo en cuenta lo anterior, resulta ser evidente que el desarrollar software para facilitar la clasificación de patologías es una tarea cada vez más necesaria, sin embargo no es fácil ejecutarla puesto que la sintomatología o atributos particulares de cada patología varían dependiendo del paciente, además actualmente existen una gran variedad de técnicas de clasificación que funcionaran mejor dependiendo de la topología de los datos o atributos recolectados de la patología. Por tanto, en este trabajo de investigación se realizará una comparación de técnicas supervisadas de clasificación multi-clase como lo son: SVM (Support Vector Machine), KNN (K-nearest neighbors algorithm), ANN (Artificial neural network), PC (Parzen´s Classifier), Random Forest o arboles de decisión

y Adaboost (Adaptative Boosting) el cual es usado en combinación con los 4 primeros clasificadores independientemente.

Método:

Materiales utilizados:

Bases de datos:

Para evaluar la metodología propuesta, se usó 3 bases de datos de UCI Machine Learning Repository. La primera, llamada cardiotocografía, contiene 2126 cardiotocogramas fetales pertenecientes a diferentes clases. Esta base de datos contiene 21 atributos que incluyen: LB-FHR Punto de referencia (latidos por minuto), AC aceleraciones por segundo, FM movimientos fetales por segundo, UC contracciones del útero por segundo, DL desaceleraciones suaves por segundo, DS Desaceleraciones severas por segundo, DP desaceleraciones prolongadas por segundo, ASTV porcentaje de tiempo con variabilidad anormal a corto plazo, MSTV promedio de variabilidad a corto plazo, ALTV porcentaje de tiempo con variabilidad anormal a largo plazo, MLTV promedio de variabilidad a largo plazo, Width: Ancho de histograma FHR, Min: Mínimo del histograma FHR, Max: Máximo del histograma FHR, Nmax: Numero de picos del histograma, Nzeros: Numero de ceros del histograma, Mode: forma del histograma, Mean: promedio del histograma, Median: mediana del histograma, Variance: Varianza del histograma, Tendency: Tendencia del histograma, CLASS FHR código de clase de la muestra (1 a 10) y NSP clase del estado fetal (Normal=1; Sospechoso=2; Patológico=3).

La segunda base de datos es llamada Cleveland, contiene 303 instancias. Contiene 13 atributos que incluyen edad, sexo, tipo de dolor de pecho, presión en descanso, colesterol, azúcar en la sangre en ayuno, ECG en descanso, ritmo cardiaco máximo, angina, pico mayor, inclinación, numero de vasos sanguíneos de color, thal y clasificación de las muestras desde 0 que significa la no presencia a 4 tipos de patologías del corazón.

La última base de datos se llama hipotiroidismo, se creó a partir de exámenes médicos reales para la detección de problemas hipotiroideos. Contiene 3772 muestras, está estructurada con 22 atributos, y 3 posibles clases que son: normal, hipotireoideo primario o hipotiroide compensado.

Clasificadores:

Los clasificadores que recomendó la literatura para realizar el estudio comparativo son : El clasificador de K vecinos más cercanos (K-NN) es un algoritmo basado en la distancia,

redes neuronales artificiales (ANN) con un enfoque basado en la búsqueda heurística, máquinas de vectores de soporte (SVM) como un clasificador basado en modelos, el clasificador Parzen (PC) es un clasificador no paramétrico basado en densidad, arboles de decisión (Random Forest) y el clasificador Adaboost, que se basa en la combinación lineal de clasificadores de manera iterativa, por lo tanto, se implementará Adaboost con cada uno de los clasificadores mencionados.

Procedimiento y Diseño:

Para empezar con el estudio comparativo se hace un pre-procesamiento a las bases de datos elegidas como se muestra a continuación:

- Pre-procesamiento:

Se realiza una selección de variables a través de los llamados "Subconjuntos de características basadas en correlación (CfsSubsetEval), algoritmo que evalúa la relevancia de un subconjunto de atributos mediante el análisis de la capacidad predictiva de cada característica junto con el grado de redundancia entre ellos. Y como método de búsqueda el algoritmo "El mejor Primero", para reducir el número de parámetros por instancia de un conjunto de datos. Como resultado de esta etapa, se obtiene que la base de datos de cardiotocografía se reduce a 10 características, hipotiroidismo a 5 atributos y la base de datos Cleveland a 7 características.

Teniendo en cuenta que los clasificadores elegidos son supervisados, es necesario realizar determinar un porcentaje de la base total para hacer el entrenamiento del clasificador. Para cada una de las bases de datos se determinó un 85% de los datos para entrenamiento de los clasificadores y un 15% de los datos para la evaluación. Es importante resaltar que la construcción de cada conjunto se hace de manera aleatoria es decir no se elige elije los datos que conformarán el conjunto de entrenamiento de manera ordenada como se encuentran en la base de datos.

- Ajuste de parámetros en los clasificadores:

Por otro lado, también se hace necesario determinar algunos atributos y métodos que rigen el buen funcionamiento de cada clasificador como se muestra a continuación:

K-NN: esta técnica de clasificación basada en la densidad de probabilidades necesita un valor para el determinar el número de vecinos (K), dicho parámetro es optimizado por medio de una estrategia de "dejar uno por uno".

ANN: La técnica de clasificación heurística requiere un número de neuronas por capa oculta. En este trabajo, se utiliza el método back-propagation en una red neuronal conocida como "alimentación hacia adelante" con una sola capa oculta. El número de neuronas en la capa oculta se calcula a partir de los propios datos, la mitad de las instancias divididas por el número de características más el número de clases. Se inicializa los pesos de la red neuronal en cero.

SVM: este método de clasificación basado en observaciones aprovecha el truco del kernel para calcular el hiperplano no lineal más discriminativo entre clases. Por lo tanto, su rendimiento depende en gran medida de la selección y ajuste del tipo de kernel. Para este estudio se selecciona un kernel gaussiano por su habilidad de generalización y su parámetro de ancho de banda ajustado por la regla de Silverman (Sheather, S. J., 2004).

Parzen: Este método de clasificación basado en probabilidades, requiere un parámetro de suavizado para la distribución Gaussiana que debe ser optimizado.

Random Forest: Para este clasificador se determina 100 árboles de decisión a generar.

Adaboost: Este método tiene varios parámetros a ajustar como se muestra a continuación:

- Clasificador: Para nuestro estudio en este parámetro se utilizaron los clasificadores SVM, ANN, Parzen y K-NN.
- Número de clasificadores: Este parámetro se refiere al número de veces que se desea ejecutar el clasificador seleccionado en el anterior parámetro.
- Regla de combinación: Se refiere al método que se utiliza para combinar los clasificadores iterados entre los que se encuentran: Promedio (Mean), votación (Vote), mínimo (Min), máximo (Max), producto (Prod), votación por pesos (Wvote) y mediana (Median). Para este trabajo se probaron todas estas reglas de combinación con cada clasificador propuesto.

- Evaluación de desempeño de los clasificadores:

Finalmente, después tener en cuenta las anteriores consideraciones se procede a determinar el desempeño de cada uno de los clasificadores, para lo cual se obtiene el error de clasificación medido de 0 a 1, donde 0 significa que el clasificador clasificó todos los datos correctamente. Para cada clasificador se realizó una validación cruzada de 20 veces para obtener el promedio de error de clasificación y su respectiva desviación estándar que

finalmente son los indicadores de desempeño a comparar presentados en la sección de resultados.

Resultados:

Aplicando la metodología propuesta en las bases de datos de Cleveland, cardiotocografía e hipotiroidismo se obtienen los resultados presentados en la tabla I y II, se obtuvo el promedio y la desviación estándar tras iterar 20 veces cada uno de los clasificadores en el caso de adaboost se varía el parámetro de combinación.

Tabla I

Error de clasificación para las bases de datos consideradas tras iterar 20 veces el clasificador Adaboost con cada combinador.

Base de Datos	Combinador	PC	KNN	SVM	ANN
CLEVELAND	Mean	0,461 ± 0,037	0,439 ± 0,055	0,466 ± 0,004	0,463 ± 0,040
	Vote	0,530 ± 0,097	0,487 ± 0,032	0,467 ± 0,000	0,502 ± 0,074
	Min	0,865 ±0,049	0,493 ± 0,082	0,629 ± 0,169	0,456 ± 0,020
	Max	0,475 ± 0,038	0,526 ± 0,101	0,466 ± 0,061	0,593 ± 0,141
	Prod	0,754 ± 0,054	0,447 ± 0,053	0,466 ± 0,010	**0,437 ± 0,048**
	Wvote	0,449 ± 0,040	0,928 ± 0,030	0,726 ± 0,173	Na
	Median	0,622 ± 0,122	0,529 ± 0,055	0,461 ± 0,000	0,546 ± 0,133
CARDIOTOCOGRAFIA	Mean	0,029 ± 0,008	0,189 ± 0,225	0,494 ± 0,292	0,138 ± 0,119
	Vote	0,080 ± 0,039	0,121 ± 0,254	0,058 ± 0,024	0,107 ± 0,120
	Min	0,201 ± 0,070	0,224 ± 0,167	0,544 ± 0,417	0,334 ± 0,266
	Max	**0,027 ± 0,007**	0,210 ± 0,167	0,715 ± 0,223	0,038 ± 0,027
	Prod	0,147 ± 0,052	0,179 ± 0,199	0,537 ± 0,431	0,256 ± 0,097
	Wvote	0,065 ± 0,036	0,187 ± 0,266	0,197 ± 0,290	Na
	Median	0,081 ± 0,032	0,189 ± 0,244	0,173 ± 0,258	0,114 ± 0,064
HIPOTIROIDISMO	Mean	**0,183 ± 0,044**	0,313 ± 0,214	0,551 ± 0,064	0,193 ± 0,082
	Vote	0,313 ± 0,080	0,312 ± 0,204	0,539 ± 0,074	0,189 ± 0,088
	Min	0,599 ±0,061	0,385 ± 0,141	0,597 ± 0,068	0,482 ± 0,146
	Max	0,183 ± 0,041	0,381 ± 0,126	0,530 ± 0,083	0,303 ± 0,033
	Prod	0,544 ± 0,095	0,309 ± 0,218	0,533 ± 0,081	0,283 ± 0,113
	Wvote	0,558 ± 0,148	0,318 ±0,218	0,568 ± 0,077	Na
	Median	0,343 ± 0,071	0,330 ±0,214	0,529 ± 0,071	0,189 ±0,096

Se puede evidenciar que la base de datos de Cleveland es un reto para los clasificadores ya que existen clases que cuentan con muy pocas muestras y esto hace que el entrenamiento no sea el adecuado, provocándose un sobre entrenamiento en las clases que tengan el mayor número de muestras, sin embargo en este tipo de bases de datos

resulta mejor el utilizar clasificadores sin ser introducidos a una estructura de replicación y ponderación de clasificadores como lo es Adaboost, lo que tiene sentido debido que este tipo de clasificadores es susceptible a overfitting, efecto de sobreentrenar un algoritmo de aprendizaje.

Tabla II

Error de clasificación tras iterar 20 veces cada base de datos con los clasificadores considerados

Clasificador	CLEVELAND	CARDIOTOCOGRAFIA	HIPOTIROIDISMO
PC	0,425 ± 0,040	0,061 ± 0,008	0,237 ± 0,041
KNN	0,416 ± 0,049	0,026 ± 0,016	0,118 ± 0,031
SVM	0,466 ± 0,004	0,022 ± 0,008	0,424 ± 0,025
ANN	0,439 ± 0,040	0,047 ± 0,023	0,179 ± 0,070
Random Forest	0,776 ± 0,071	0,395 ± 0,087	0,308 ± 0,152

Para bases de datos que tienen mejor equilibrado su número de muestras para cada clase como cardiotocografía se obtienen mejores resultados al ser entrenado con Adaboost ya que al ser de tipo adaptativo puede mejorar las instancias mal clasificadas por los clasificadores anteriores.

En cuanto a hipotiroidismo ocurre lo mismo que para Cleveland ya que esta también cuenta con clases que tienen pocas muestras, a pesar de eso los clasificadores independientes como SVM y KNN funcionan muy bien, se obtiene un error de clasificación bajo lo que garantiza un mejor diagnóstico.

Conclusiones:

De los resultados arrojados para las pruebas realizadas, resulta claro el evidenciar que el algoritmo de RandomForest presenta un error de clasificación bastante alto en comparación con los demás clasificadores usados en el experimento. Por lo tanto no es el clasificador más adecuado para ser adaptado en algoritmos que tomen decisiones de diagnóstico de patologías como por ejemplo estructuras como razonamiento basado en casos, CBR por sus siglas en ingles. Lo anterior se puede deber quizá a la topología de los datos dependiendo de esta, existen clasificadores que funcionan de mejor manera, es importante el reconocer esto ya que podría facilitar tareas. Con el hecho de solo saber la topología de los datos se podría predecir que clasificador funcionará mejor.

Basándose en los resultados obtenidos se concluye que el desempeño de un clasificador va a depender en gran medida de la naturaleza de los datos y de la cantidad de muestras

que contenga la base de datos. Resulta muy interesante que en algunos casos clasificadores sencillos tienen mejores resultados con respecto a clasificadores más sofisticados.

Como Parte de trabajos futuros se recomienda tener en cuenta las pruebas realizadas para que los clasificadores con mayor rendimiento sean utilizados en estructuras de toma de decisiones y diagnostico como lo son CBR, además se podría realizar balanceo de clases en aquellas bases de datos que cuentan con pocas muestras de cierta clase esto evitaría el sobre-entrenamiento y mejoraría el rendimiento de los clasificadores.

Referencias:

Reyes, S. O. L., Colín, G. M., & Murata, C. (2014). Inteligencia artificial para asistir el diagnóstico clínico en medicina. Revista Alergia México, 61(2).

Sheather, S. J. (2004). Density estimation. Statistical Science, 19(4), 588-597.

Siegel, E. (2013). Predictive analytics: The power to predict who will click, buy, lie, or die. John Wiley & Sons.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

.