Análisis del desempeño de algoritmos para la construcción de árboles de decisión.

Trabajo terminal No. _____

Alumnos: **Blancas Analco Raúl Ajaib*
Directores: *Miriam Pescador Rojas, Víctor Adrián Sosa Hernández.*
*e-mail: *rblancasa0900@alumno.ipn.mx*

Resumen. Propuesta de desarrollo de trabajo terminal para la implementación y comparación de algoritmos genéticos, de inducción e híbridos para la construcción de árboles de decisión para problemas de clasificación.

Palabras clave: inteligencia artificial, aprendizaje automático, algoritmo genético, árbol de decisión, clasificador estadístico.

1. Introducción.

En análisis de operaciones e informática estadística cuando un sistema de información es diseñado y construido con el objetivo de brindar soporte a un proceso de toma de decisiones sobre conjuntos de datos cambiantes y estructurados recibe el término de Sistema de Soporte de Decisiones (*Decision Support System, DSS*). Un DSS está compuesto por tres elementos generales, primeramente la fuente de datos que proveerá la información a ser procesada, pudiendo ser ésta de naturaleza variable: una base de datos, un conjunto de archivos indexados, información tabular estructurada, etc.; el segundo elemento es el modelo dado a los datos obtenidos de una o más fuentes de información, entiéndase esto como el contexto de uso y criterios de decisión a aplicarse sobre la información posterior a procedimientos de limpieza y formato estructurado; el tercer elemento corresponde a la manera de representar el conocimiento obtenido tras la aplicación del modelo, dependiendo esto ampliamente de las características de los usuarios que consumen la información generada, destacan por su amplio uso la visualización tabular interactiva y los gráficos visuales de dos o tres dimensiones [1].

Es destacable el papel que toma el modelo de datos en un DSS, el modelo es el elemento responsable de definir el comportamiento que el sistema tendrá y la naturaleza de los resultados por obtener. Entre los posibles comportamientos a tomar por un DSS son: sistemas de comparación de variables o conjuntos de datos diferentes, sistemas estadísticos predictivos que permiten inferir comportamientos futuros de las variables analizadas o sistemas estadísticos descriptivos que especifican el estado de los datos mediante la agrupación o categorización de estos o explican relaciones implícitas entre ellos mediante tareas de hallazgo de dependencias o clasificación.
La estadística descriptiva como elemento de clasificación cumple con el objetivo de identificar a qué categoría, grupo o clase pertenece un conjunto de rasgos o características que describen las entradas que conforman un conjunto de datos estructurados.
Gran variedad de técnicas se utilizan para llevar a cabo un proceso de clasificación, destacando por su amplio uso: regresiones logísticas la cual hace uso de una función sigmoide para modelar una variable dependiente binaria, máquinas de soporte vectorial que delimita un espacio geométrico trazando un camino que sirve como clasificador, redes neuronales que hace uso de transformaciones lineales, trigonométricas y de dimensionalidad como método de clasificación o árboles de decisión basado en reglas de decisión lógicas [2].

Dentro de la gama de herramientas con capacidad para realizar un clasificador, la simplificación de interpretación y flexibilidad para el entendimiento y análisis multidisciplinario provisto por los árboles de decisión los convierten en una opción empresarial y académica de uso generalizado sin sacrificar la exactitud y eficiencia de los resultados que proveen.
Un árbol de decisión es una herramienta que se describe como un conjunto de vértices o nodos interconectados por aristas o enlaces con la particularidad de que para cada par de nodos existe uno y sólo un camino entre ellos, equivalentemente, puede interpretarse como un grafo conexo acíclico no dirigido; donde las 'ramas' representan limitaciones u observaciones de las características de los individuos de estudio, hecho esto mediante la definición de reglas lógicas de decisión, y las 'hojas' representan la clase o etiqueta dentro de los valores objetivo que poseen los individuos de estudio, los valores en las hojas se asignan siguiendo las reglas definidas en las ramas que conforman el camino desde un nodo raíz o rama principal hasta la hoja. Cuando los valores asignados a las 'hojas' del árbol toman valores meramente discretos, el árbol de decisión funciona para propósitos de clasificación [3].
El objetivo principal de un árbol de decisión es crear un modelo capaz de predecir el valor de una variable objetivo basado en los valores particulares que tomen el conjunto de variables de entradas. En un árbol de decisión a cada nodo interno (nodo no hoja) se le asigna un rasgo definido por una sola variable de entrada; a los arcos que conectan un nodo interno se les asignan los posibles valores que toma la variable de

entrada asociada a él, cada arco conecta al nodo interno con otro nodo interno o con un nodo hoja; a cada nodo hoja se le asigna una etiqueta o clase, un valor particular dentro del rango de valores de la variable objetivo.
El método general para construir un árbol de decisión consiste en dividir progresivamente el conjunto de factores o variables de entrada en subconjuntos de menor dimensión, en cada división se genera una regla sobre la cual se toma una decisión sobre un factor de entrada particular, las reglas generadas en conjunto definen la clasificación a obtener. La división en subconjuntos de las variables de entrada es realizada recursivamente hasta que el nodo en proceso actual cuente con todos sus elementos dentro del mismo valor en la variable objetivo o la división propuesta no añade valor alguno en la clasificación. El proceso de división en subconjuntos recursivamente es un algoritmo heurístico de carácter voraz [4].

Alternativamente, un algoritmo genético es una técnica computacional utilizada para generar soluciones de alto grado de aceptación a problemas de búsqueda y optimización usando como marco de acción el proceso biológico de selección natural. En un algoritmo genético un conjunto de soluciones candidatas (usualmente llamados individuos) al problema de optimización a resolver sigue un proceso evolutivo progresivo hacía una o múltiples soluciones finales [5].
El proceso de evolución progresiva inicia con una población de individuos generada aleatoriamente llamada generación inicial, cada individuo es sometido a una evaluación de aptitud, la cual corresponde al valor de la función objetivo en el problema de optimización a resolver. Con base en los conceptos de selección natural de las especies, los individuos con mayor aptitud (las soluciones mejor evaluadas en la función objetivo) dentro de la generación en proceso son seleccionados estocásticamente y se les aplica operadores genéticos, de cruce o reproducción y mutación o alteración, para formar una nueva generación la cual es usada en la siguiente iteración del algoritmo. Este proceso se repite un número finito de generaciones que es definido al inicio del proceso y al finalizar se reportan las mejores soluciones obtenidas de acuerdo a su aptitud [5].

En un algoritmo genético estándar son requeridos los siguientes elementos para su ejecución: representación genética, evaluación de aptitud, población inicial, selección, operadores genéticos de cruce y mutación, así como un criterio de terminación [5].

En el presente protocolo y posterior trabajo terminal se pretende el conjuntar las dos técnicas previamente descritas, los árboles de decisión y algoritmos genéticos sobre un mismo marco de trabajo que comparará tres enfoques de construcción a partir de su desempeño como herramienta de clasificación.

La tabla 1 muestra un resumen de implementaciones de uso libre de algoritmos de aprendizaje de máquina con relación al trabajo terminal que se propone.

Software	Características	Precio en el mercado
Weka [6]	Colección de algoritmos y herramientas de visualización para minería de datos y modelado predictivo. Diversos algoritmos de árboles de decisión y optimizaciones particulares por algoritmo.	Software libre sobre Licencia General Pública GNU.
KNIME [7]	Plataforma para la analítica de datos, creación de reportes e integración de repositorios de datos. Implementa múltiples técnicas de aprendizaje máquina y minería de datos.	Software libre sobre Licencia General Pública GNU.
Scikit-learn [8]	Biblioteca de aprendizaje de máquina en lenguaje de programación Python.	Software libre sobre Licencia New BSD.
Chefboost [15]	Biblioteca que implementa un conjunto de algoritmos de árbol de decisión tradicionales y avanzados en lenguaje de programación Python.	Software libre bajo Licencia MIT.

Tabla 1. Resumen de productos similares.

Por otro lado, la tabla 2 muestra un resumen de trabajos de investigación de algoritmos genéticos híbridos, lo cual tiene una alta relación con la presente propuesta de trabajo terminal.

Artículos de investigación	Características
A Hybrid Decision Tree / Genetic Algorithm Method for Data Mining. [9]	Propone el uso de un algoritmo genético para optimizar el conjunto de reglas de decisión lógicas generadas por un algoritmo de árbol de decisión cuya frecuencia de la variable objetivo sea de un aporte considerablemente menor a sus predecesores.
Decision Tree Classifier for Network Intrusion Detection with GA-based Feature Selection. [10]	El banco de origen de datos usado en este artículo posee una alta dimensionalidad y el algoritmo genético es usado para elegir los factores o variables de entrada para componer el árbol de decisión, este último es generado con un algoritmo de árboles de decisión.

Tabla 2. Resumen de documentos similares.

2. Objetivo.

Analizar y comparar el desempeño obtenido por tres algoritmos diferentes para la construcción de árboles de decisión enfocados en la clasificación: un algoritmo genético, un método tradicional de inducción (C4.5) y un método híbrido (tradicional - algoritmo genético).

Objetivos particulares:

- Selección de orígenes de bases de datos de trabajo.- buscar y seleccionar bases de datos para clasificación con diferentes características.
- Implementación de un algoritmo genético para la clasificación.- implementar un algoritmo genético cuya salida corresponda a un conjunto de reglas lógicas de decisión, desarrollando para ello los siguientes elementos: representación de individuos, evaluación de desempeño por regla de split candidato *Twoing*, operadores genéticos de selección, cruza, mutación y criterio de paro.
- Adaptación de un híbrido de árbol de decisión y algoritmo genético .- ejecutar un algoritmo de árbol de decisión, posteriormente seleccionar un subconjunto de reglas encargadas de evaluar una cantidad de registros del origen de datos menor a un porcentaje por definir en la fase de experimentación que será sometido a un proceso de optimización por medio de un algoritmo genético.
- Ejecución de un algoritmo inducción tradicional.- obtener resultados de ejecución del algoritmo en los orígenes de datos seleccionados.
- Evaluación de algoritmos con orígenes de datos.- evaluar el desempeño de los algoritmos con indicadores apropiados.
- Comparativa de resultados.- comparar los resultados obtenidos para proveer información sobre qué método es adecuado o recomendado para las diferentes características en los posibles orígenes de datos.

3. Justificación.

Los algoritmos estándares usados para la construcción de árboles de decisión basan su comportamiento en la división del conjunto de variables de entrada progresivamente de manera recursiva, en cada división se elige entre el conjunto de variables disponibles una sóla variable y se genera una regla de decisión lógica sobre los valores en el rango de dicha variable, para la elección de una variable de decisión entre las variables de entrada disponibles se evalúan dichas variables a través de indicadores cuyo objetivo es evaluar que tan bien o que tan mal se ha dividido la información que se está analizando utilizando funciones evaluadoras: entropía, ganancia de información, índice de gini, entre otras [11].

Características destacables de los árboles de decisión:

- En comparación con otros mecanismos usados para la clasificación son capaces de manejar datos tanto numéricos como categóricos.
- Usan un modelo de "caja blanca", es decir las condiciones que genera el modelo de procesamiento pueden ser fácilmente interpretadas usando lógica booleana; es altamente interpretable bajo un criterio de decisiones humanas.
- Los árboles de decisión no son lineales, permiten la subdivisión de la información bajo modelos no paramétricos.

La capacidad de los algoritmos genéticos de realizar búsquedas sobre toda la amplitud del rango de acción definido en su genoma mediante la inicialización de una población lo suficientemente amplia y diversa apoyándose de mecanismos de aleatoriedad, le permiten un alto desempeño en la resolución de problemas de optimización global.

Los operadores genéticos de cruce y mutación están diseñados para lograr que los procesos de búsqueda se alejen de los óptimos locales en contraste con algoritmos de aproximación tradicionales. En contraste a la búsqueda de soluciones robustas con un alto rendimiento probando todas las combinaciones del rango de búsqueda de los datos provistos, se construyen soluciones sencillas paulatinamente mejor adecuadas a lo óptimo de su aptitud a partir de soluciones previas, en este caso particular se le da el nombre de generaciones. Al representar un individuo como una serie o conjunto de propiedades, estas propiedades reciben el nombre de cromosomas o alelos, se brinda la libertad de realizar operaciones cortas y de bajo orden que al recombinarse y cambiar levemente incrementan potencialmente los óptimos de aptitud buscados. [12]

Al aprovechar la sencillez de la interpretación tan humanamente cercana que ofrecen los árboles de decisión en conjunto con la capacidad de los algoritmos genéticos para realizar búsquedas de carácter global o usarse como un elemento de optimización adicional, un modelo de clasificación robusto y de fácil interpretación es obtenido, sin embargo no existe un estudio que analice o compare el comportamiento de un algoritmo genético con capacidades de generar un árbol de decisión con alguna otra técnica de clasificación.
En el presente trabajo terminal se pretende validar el desempeño de dicho modelo de clasificación mediante un proceso iterativo e incremental de experimentación y comparación usando como medidas de desempeño exactitud y área bajo la curva.

Dentro de las vertientes de un problema de investigación se tienen tres ejes de acción principales: dar resolución a un problema específico, probar una teoría o aportar evidencia empírica en favor de ella y generar el planteamiento a un problema o inducir el conocimiento [13]. El presente protocolo de trabajo terminal corresponde a la tercera vertiente de un trabajo de investigación, en particular la propuesta de comparativa de tres enfoques diferentes de árbol de decisión basados en los indicadores de desempeño de cada enfoque a evaluar.

4. Productos o resultados esperados.

De acuerdo a los objetivos planteados, se enlistan los entregables contemplados tras la finalización del presente protocolo como Trabajo Terminal:

- Informe, bitácora de experimentos y evaluación de desempeño.- resultados de las evaluaciones de los tres métodos a analizar: algoritmo genético, árbol de decisión con optimización por algoritmo genético y árbol de decisión por algoritmo tradicional, usando los siguientes indicadores: Índices de exactitud, índice de área bajo la curva, validación por k-fold cross.
- Comparativa estadística de resultados obtenidos.- comparativa de los tres métodos usados usando tabulados o gráficos para la mejor visualización de los índices de desempeño e informe de aplicación recomendada según las características de los orígenes de datos usados.
- Wiki del proyecto.- Reporte y guías o manuales de usuarios requeridos como parte del proceso de la asignatura trabajo terminal.
- Reporte técnico de trabajo terminal.

5. Metodología.

Para el presente Trabajo Terminal se propone dada su naturaleza de aporte al conocimiento más que aplicación técnica o resolución de problemáticas particulares, el uso de una metodología científica de investigación basada en pruebas para su desarrollo, entiéndase dicha metodología como un proceso iterativo y potencialmente incremental que contempla las siguientes fases: hipótesis, construcción de elementos de evaluación, pruebas y experimentación, obtención e interpretación de resultados y ajustes o modificaciones del marco de trabajo planteado [14]. Se subdivide el proceso en cuatro etapas principales:

- Etapa cero.- delimitación de hipótesis y marco de trabajo, investigación de técnicas y metodologías y selección de bases de datos.
- Etapa uno.- construcción de un algoritmo genético que genere reglas lógicas de decisión, experimentación con orígenes de datos variados y evaluación de resultados.
- Etapa dos.- ejecución de algoritmo de árbol de decisión, evaluación y selección de porcentaje de reglas generadas por frecuencia de registros dada la variable objetivo clasificada, optimización de reglas seleccionadas por algoritmo genético, experimentación con orígenes de datos variados y evaluación de resultados.
- Etapa tres.- ejecución de algoritmo de construcción de árbol de decisión tradicional, experimentación con orígenes de datos variados y evaluación de resultados.
- Etapa cuatro: comparativa estadística de resultados obtenidos en las etapas anteriores.

6. Cronograma.

Actividad	AGO	SEP	OCT	NOV	DIC	FEB	MAR	ABR	MAY	JUN
Etapa cero										
Delimitación de hipótesis y marco de trabajo inicial.										
Investigación de metodologías, técnicas y algoritmos.										
Selección de bases de datos diversas.										
Etapa uno										
Representación y población inicial.										
Operadores genéticos.										
Operadores genéticos.										
Experimentación										
Ajuste de parámetros del algoritmo.										
Evaluación										
Actualización de documento técnico										
Etapa dos										

Ejecutar árbol de decisión.										
Selección de porcentaje de reglas de baja dimensión										
Optimización de reglas por algoritmo genético.										
Experimentación										
Evaluación										
Actualización de documento técnico										
Presentación Trabajo Terminal I										
Evaluación de comentarios de presentación.										
Etapa tres										
Ejecución árbol de decisión.										
Evaluación										
Experimentación										
Actualización de documento técnico										
Etapa cuatro										
Ajuste de parámetros del algoritmo.										
Comparativa estadística.										
Análisis de resultados.										
Wiki del proyecto.										
Actualización de documento técnico										
Presentación Trabajo Terminal II										
Evaluación de comentarios de presentación.										

7. Referencias.

[1] G. M. Marakas, *Decision Support Systems in the Twenty-first Century*, Prentice Hall, 1999.
[2] C. C. Aggarwal, *Data Classification: Algorithms and Applications*, CRC Press, 2014.
[3] S. S. Shwartz, *Understanding Machine Learning from Theory to Algorithms*, Cambridge University Press, 2014
[4] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, 2 edición, Springer, 2009.

[5] S.N. Sivanandam, S.N. Deepa, *Introduction to Genetic Algorithms*, Springer, 2008.
[6] University of Waikato, "Machine Learning at Waikato University", cs.waikato.ac.nz, [Online] Available: https://www.cs.waikato.ac.nz/ml/index.html
[7] KNIME AG, "KNIME Software Overview", knime.com, [Online] Available: https://www.knime.com/software-overview
[8] F. Pedregosa, *Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research*, 2011.
[9] D. R. Carvalho, A. A. Freitas, *A Hybrid Decision Tree / Genetic Algorithm Method for Data Mining*, Proceedings of the Genetic and Evolutionary Computation Conference, USA, 2004.
[10] A. S. Wu, K. A. Hua, *Decision Tree Classifier for Network Intrusion Detection with GA-based Feature Selection*, ACM-SE Conference, USA, 2005.
[11] L. Breiman, J. Friedman, *Classification and Regression Trees*, CRC Press, 1 digital edition, 2017.
[12] Cruz-Meza M.E., Curso de Algoritmos Genéticos (2019 - 2020 A), Genetic Algorithms C374, México: ESCOM - IPN.
[13] R. Sampieri, *Metodología de la investigación*, 6ta edición, McGraw-Hill, 2014.
[14] G. Baena-Paz, *Metodología de la Investigación: Serie Integral por Competencias*, 3ra edición, Patria, 2017.
[15] S. I. Serengil, ChefBoost: *A Lightweight Boosted Decision Tree Framework*, Zenodo, 2021.

8. Alumno y Directores.

CARÁCTER: Confidencial
FUNDAMENTO LEGAL: Artículo 11 Fracc. V y Artículos 108, 113 y 117 de la Ley Federal de Transparencia y Acceso a la Información Pública.
PARTES CONFIDENCIALES: Número de boleta y teléfono.

Raúl Ajaib Blancas Analco. Alumno de la carrera de Ing. en Sistemas Computacionales en ESCOM, Boleta: 2014630050, Tel. 5527714840, email. sayaann95@gmail.com

Firma: _____

Pescador Rojas Miriam.

Dra. en Ciencias en Computación por el CINVESTAV-IPN en 2019, M. en C. en Computación por el CINVESTAV-IPN en 2010, Ing. en Sistemas Computacionales por la ESCOM-IPN en 2008, Profesora de carrera en ESCOM-IPN desde 2010 a la fecha, áreas de interés Inteligencia Artificial, Cómputo Evolutivo, Aprendizaje Máquina.
Tel: 57296000. Ext: 52022.
E-mail: mpescadorr@ipn.mx

Firma: _____

Sosa Hernández Víctor Adrián.

Dr. en Ciencias en Computación por el CINVESTAV-IPN en 2017, M. en C. en Computación por el CINVESTAV-IPN en 2013, Ing. en Sistemas Computacionales por la ESCOM-IPN en 2011, Profesor del Departamento de computación en la Escuela de Ingeniería y Ciencias del ITESM campus Edo. de México desde el 2017 a la fecha, áreas de interés Inteligencia Artificial, Cómputo Evolutivo, Aprendizaje Máquina, Cómputo en la nube.
E-mail: vsosa@tec.mx

Firma: _____

[ADVERTENCIA, MENSAJE EXTERNO] Re: Versión entregable con correcciones de protocolo de trabajo terminal - Blancas Analco

1

R

Raul Ajaib Blancas Analco
Vie 29/04/2022 08:20 AM

Para: Miriam Pescador Rojas; Víctor Adrián Sosa Hernández

Protocolo_AD-AG.pdf
153 KB

Buen día Dra. Miriam Pescador y Dr. Victor Sosa,

Hago envió de protocolo de trabajo terminal titulado "Análisis del desempeño de algoritmos para la construcción de árboles de decisión" con calendario ajustado a dos periodos escolares y actualizaciones en el titulo y ajuste de parámetros para su confirmación de recepción y aceptación de contenido.

Atentamente,
Blancas Analco Raúl Ajaib
Boleta: 2014630050

Miriam Pescador Rojas
Vie 29/04/2022 08:57 AM

Para: Raul Ajaib Blancas Analco; Víctor Adrián Sosa Hernández

Buen día, acuso de recibido y doy mi visto bueno del documento de protocolo.

Saludos cordiales.

...

V

Victor Adrian Sosa Hernandez <vsosa@tec.mx>
Vie 29/04/2022 07:03 PM

Para: Miriam Pescador Rojas
CC: Raul Ajaib Blancas Analco

ADVERTENCIA, REMITENTE EXTERNO
ESTE MENSAJE SE ORIGINÓ FUERA DE LOS SERVICIOS INSTITUCIONALES. NO HAGA CLIC EN ENLACES NI ABRA ARCHIVOS ADJUNTOS Y LO MAS IMPORTANTE NO PROPORCIONE INFORMACIÓN A MENOS QUE RECONOZCA AL REMITENTE Y TENGA CERTEZA QUE EL CONTENIDO ES SEGURO.

Buen día,

doy acuse de recibido y doy mi visto bueno del documento de protocolo.

Saludos

...

Responder | Responder a todos | Reenviar