INVESTIGACIÓN

Comparación de Ténicas de Clasificación de Datos, Aplicado al Problema de la Desertificación de Suelos

Sandra Patricia Castillo L[1] , Pablo Eduardo Caicedo R.[1] y María Isabel Velandia C.[2]
1 Grupo Investigación. GRINAU, Corporación Universitaria Autónoma del Cauca, Popayán, Colombia.
2 Proyecto SIMCI, Oficina contra las Drogas y el Delito de las Naciones Unidas, Bogotá, Colombia .

Recibido: 2 de junio de 2011; revisado: 26 de junio de 2011; aceptado: 19 de julio de 2011.

Resumen— La desertificación de los suelos es considerado un problema de desarrollo sostenible relacionado con la pobreza presente en el ambito rural, además de instar otros conflictos a nivel social y económico, de seguridad alimenticia, desplazamientos, entre muchos otros.

Dada su relevancia en el orden nacional e internacional, es importante para las autoridades ambientales conocer la incidencia de algunas variables en el grado de desertificación del suelo. En este trabajo se emplearon técnicas y herramientas de minería de datos para extraer estos patrones de conocimiento a través del análisis de un conjunto de datos.

En la primera parte, se presenta el concepto de desertificación, causas e impactos, enfatizando en cómo este problema ambiental esta afectando a Colombia. Posteriormente se detalla la metodología empleada.

Finalmente se presentan resultados referidos al uso de diferentes técnicas de clasificación y desempeño de herramientas de software para comprobar el poder predictivo de algunas variables básicas incidentes (Tipo de Suelo, Erosión Hídrica, Erosión Eólica, Deforestación y Región Climática) en el grado de desertificación de un suelo.

Palabras Clave: Desertificación, Minería de Datos, Clasificación, Árboles de Decisión, Redes Neuronales Artificiales.

Abstract— The soil desertification is considered a sustainable development issue, related to rural poverty, in addition it generates social and economical problems, food safety problems, between others.

Considering his relevancy in the national and international order, it is important for the environmental authorities to know the incidence of some variables in the desertification grade of the soil. This paper shows the use of data minning techniques for the knowledge pattern extraction through information dataset analisys.

The first part of the paper shows the desertification concept, cause and impact, it emphasizes how is the affection in Colombia. Later the paper details the work's methodology.

Finally it shows the results refered to the use of different classification techniques and the performance of the software tools for the comprobation of the predictive power of the basic incident variables (Soil Type, Hydric Erosion, Eolic Erosion, Deforestation, and Regional Weather) in the desertification grade of the soil.

Keywords: Desertification, Data Mining, Classification, Decision Trees, Artificial Neural Networks.

I. INTRODUCCIÓN

Un suelo degradado se caracteriza porque su capacidad para generar biomasa ha disminuido, es decir, el potencial para albergar vida se ve afectado por diversas causas, siendo las más representativas las variaciones climáticas y el inadecuado uso humano. Todo lo anterior ocasiona la disminución en la productividad de los suelos.

Este fenómeno se puede manifestar en cualquier ecosistema, pero cuando afecta zonas secas se cataloga como desertificación, y es considerada una situación de mayor gravedad por las características del medio: suelos frágiles, vegetación escasa y el clima seco. De acuerdo CON las estadísticas, se considera que en el mundo cerca del 70% de las tierras secas dedicadas a la agricultura se encuentran arruinadas. Para el caso colombiano, según el Programa de las Naciones Unidas para el Desarrollo en Colombia (PNUD) el 17% del territorio presenta rastros de desertificación y otro porcentaje considerable es vulnerable de ser afectado a futuro.

II. TEORÍA

La Convención de las Naciones Unidas de lucha contra la Desertificación y la Sequía (UNCCD) define la desertificación como el proceso de "degradación de las tierras en las zonas áridas, semiáridas y subhúmedas secas resultante de diversos factores tales como las variaciones climáticas y las actividades humanas. La desertificación es un proceso dinámico que se observa en los ecosistemas secos y frágiles. Incluye áreas terrestres (suelo, subsuelo, acuíferos), poblaciones animales y vegetales, y los establecimientos humanos y sus servicios (como terrazas y represas, por ejemplo)."[1]

A. Causas y consecuencias de la desertificación[1][2][3]

La asociación de complejos agentes físicos, biológicos, políticos, sociales, culturales y económicos crean las condiciones que conducen a la desertificación. Otros factores considerados son:

- ***Las variaciones climáticas***: altas temperaturas durante periodos de tiempo prolongados que generan sequías y afectan el crecimiento de la vegetación.
- ***Las actividades humanas*** relacionadas principalmente con la agricultura: sobrepastoreo, cultivo excesivo,

deforestación e inadecuadas prácticas humanas.

La desertificación tiene graves efectos físicos y biológicos sobre el ambiente, lo cual termina por afectar la calidad de vida del hombre; por mencionar algunos se tienen: erosión o desaparición del suelo, escasez del agua, pérdida de biodiversidad, aumento de la pobreza y la migración de la población.

Son pocas las áreas que exhiben un proceso natural de desertificación, sin embargo por acción antrópica, zonas áridas y semiáridas pueden convertirse en nuevos desiertos (ver Fig. 1); se habla de siete procesos principales de desertificación [6]:

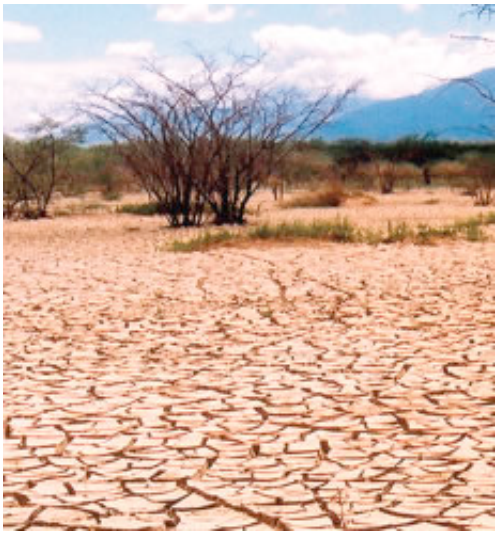

Fig. 1. Aspecto de un suelo desertificado. Fuente [5]

- ***Daño de la cubierta vegetal:*** deforestación producida por la tala, los incendios, la lluvia ácida, etc.
- ***Erosión hídrica:*** se genera cuando las corrientes de agua devastan la superficie que cubre el suelo, incluye procesos como la erosión por salpicadura, la erosión laminar, la erosión en cárcavas, entre otros.
- ***Erosión eólica:*** eliminación de la cubierta del suelo causada por el viento.
- ***Exceso de sales:*** causado por la alta concentración de sal en el suelo, genera una reducción considerable en el desarrollo vegetal.
- ***Degradación biológica:*** generada cuando es arrancado el suelo vegetal que suministra los nutrientes orgánicos al suelo.
- ***Degradación física:*** cambios hostiles en las propiedades físicas de suelo (porosidad, permeabilidad, densidad aparente, estabilidad estructural, etc).
- ***Degradación química:*** provocada por procesos como la lixiviación de bases y la formación de diferentes áreas de toxicidad en el suelo, debidas al exceso de sales.

B. El problema de la desertificación en Colombia [4]

De acuerdo con las condiciones topográficas del país, los procesos de ocupación-apropiación del territorio y los modelos de consumo-producción, han ocasionado que la degradación de los suelos se acrecente conllevando a una modificación de los ecosistemas originarios; la mayor afectación se encuentra en los bosques húmedos tropicales, los bosques secos, los bosques andinos, los páramos, las sabanas del Caribe y Orinoquia y los ecosistemas de manglar (ver Fig. 2). En cifras, se considera que el 16.95% del territorio nacional está afectado por el fenómeno de desertificación y en el 78.9% de las zonas secas se manifiestan diferentes niveles de desertificación.

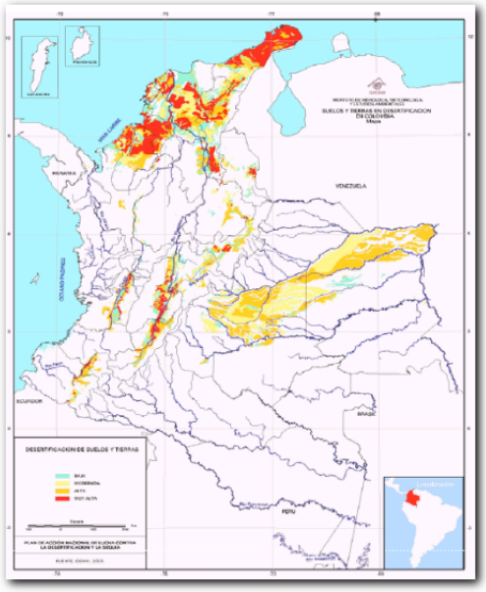

Fig. 2. Mapa de desertificación en Colombia. Fuente IDEAM

III. DEFINICIÓN DEL PROBLEMA

El trabajo se orientó en la comparación de los resultados arrojados por diferentes tipos de clasificadores (árboles de decisión y redes neuronales) que buscaron estimar el grado de desertificación de un suelo a partir de un conjunto de variables básicas incidentes (Tipo de Suelo, Erosión Hídrica, Erosión Eólica, Deforestación y Región Climática).Adicionalmente se contrastó el desempeño de las herramientas de software utilizadas.

En minería de datos, la clasificación supervisada de datos tiene como objetivo asignar registros a una clase de la manera más precisa. Para cumplir con este propósito se llevan a cabo dos procedimientos: primero, se construye un modelo a partir de unos datos de ejemplo, donde cada registro pertenece a una clase específica conocida; segundo, el modelo se usa para asignar nueva información a las clases previamente establecidas[7]. Existen diferentes métodos para realizar una clasificación:

- Inducción de árboles de decisión
- Redes Neuronales Artificiales
- Métodos Bayesianos
- Clasificación basada en asociación
- Vecino más cercano

- Razonamiento Basado en Casos
- Algoritmos Genéticos
- Conjuntos Difusos

Para evaluar y comparar los diferentes métodos de clasificación se tienen en cuenta varios aspectos, entre los que se encuentran:

- ***Exactitud de predicción:*** habilidad del modelo de predecir correctamente la etiqueta de clase de nuevos ejemplos.
- ***Velocidad:*** tiempo para construir el modelo y el tiempo para usar el modelo.
- ***Robustez:*** manejo de valores faltantes y ruido (predicciones correctas).
- ***Escalabilidad:*** eficiencia en grandes bases de datos y la facilidad de interpretación. También se evalúa el nivel de entendimiento provisto por el modelo.
- ***Forma de las reglas:***
  - Tamaño del árbol de decisión
  - Qué tan compactas son las reglas de clasificación

IV. PROCEDIMIENTO

A. *Software utilizado*

Para el desarrollo del trabajo se empleó UDMiner para el pre-procesamiento de los datos, Weka fue usada para construir modelos de árboles de decisión y redes neuronales, en MatLab se utilizó Neural Networks Toolbox para construir, entrenar y probar redes neuronales, usando diferentes funciones de activación.

B. *Variables y reglas*

Para predecir la desertificación de los suelos, inicialmente fueron consideradas cinco variables: Tipo de suelo, Erosión Hídrica, Erosión Eólica, Deforestación y Región Climática, cada una con sus respectivas hipótesis, como se muestra en la Tabla 1; como variable de salida se obtiene el grado de desertificación de un suelo (Muy baja, Baja, Media, Alta o Muy alta).

C. *Análisis y preparación de los datos*

Se procesaron un total de 1.500 registros, que corresponden a todas las posibles combinaciones de los atributos. Con el fin de reducir el ruido en los datos, se utilizó la herramienta UDMiner para hacer una selección de atributos, usando una selección del 30% y una poda de 5%. Este proceso mostró como resultado que la variable Tipo de suelo no estaba directamente relacionada con el proceso de desertificación, razón por la cual no fue tenida en cuenta en la construcción de los modelos.

TABLA I
VARIABLES PARA ESTIMAR EL GRADO DE DESERTIFICACIÓN DE UN SUELO

Tipo de suelo	Erosión hídrica	Erosión eólica	Deforesta-ción	Región climática
Franco gravilloso-franco arenoso	Muy baja	Muy baja	Muy baja	Árida
Franco arenoso - franco limoarenoso	Baja	Baja	Baja	Semiárida
Franco arcilloso	Media	Media	Media	Subhúmeda seca
Arcilloso	Alta	Alta	Alta	
	Muy alta	Muy alta	Muy alta	

Para construir las redes neuronales, los atributos fueron representados usando valores numéricos como se aprecia en la Tabla 2.

TABLA II
REPRESENTACIÓN NUMÉRICA DE LOS ATRIBUTOS

Erosión hídrica, erosión eólica, deforestación	Valor	Región climática	Valor
Muy baja	1	Subhúmeda secas	1
Baja	2	Semiárida	2
Media	3	Árida	3
Alta	4		
Muy alta	5		

D. *Construcción de modelos y análisis de resultados*

A continuación se describen los detalles de cada uno de los modelos construidos usando las herramientas Weka y MatLab.

1. Procesamiento de datos en Weka

El árbol de decisión obtenido es bastante denso, tiene 235 niveles y el tamaño es de 294, como se aprecia en la Fig. 3.

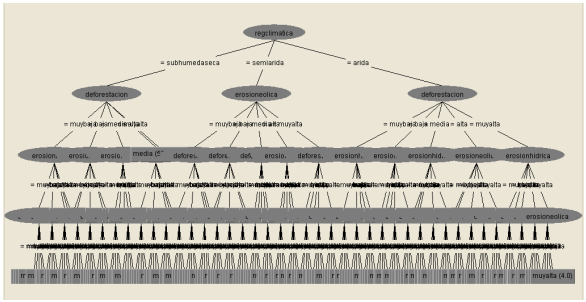

Fig. 3. Árbol de decisión obtenido con Weka

Los resultados conseguidos durante el entrenamiento son bastante buenos.

Instancias correctamente clasificadas	*1046*	*99,619 %*
Instancias incorrectamente clasificadas	*4*	*0,381 %*
Error medio absoluto	*0,0028*	
Error medio cuadrático	*0,0375*	
Error absoluto relativo	*1,0452 %*	
Error cuadrático relativo	*10,2283 %*	

Sin embargo, en el proceso de evaluación se obtuvieron resultados bastante regulares, que hacen pensar que los datos tiene un alto porcentaje de desorden.

Instancias correctamente clasificadas	*234*	*52 %*
Instancias incorrectamente clasificadas	*216*	*48 %*
Error medio absoluto	*0.205*	
Error medio cuadrático	*0.4142*	
Error absoluto relativo	*74.6889 %*	
Error cuadrático relativo	*110.7741 %*	

Para el caso de las redes neuronales, se utilizó el Perceptrón Multicapa con diferentes opciones. En primer lugar, se usaron las opciones por defecto del programa (el número de capas ocultas es automático) y el número de épocas fue 1.000. Cabe anotar que la función de activación utilizada por esta herramienta es la Sigmoide. Gráficamente la red se muestra en la Fig. 4

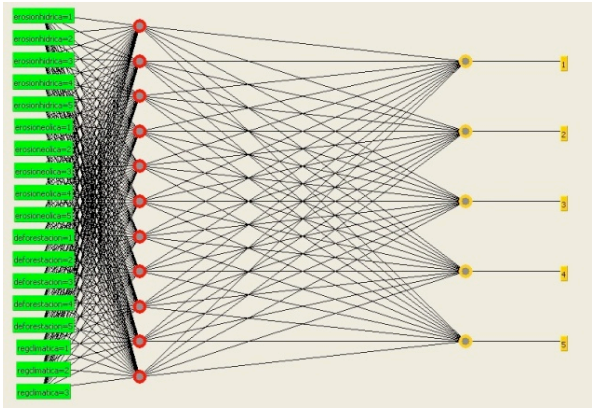

Fig. 4. Red neuronal para el problema de desertificación de suelos usando los parámetros por defecto de Weka

Los resultados alcanzados para el proceso de entrenamiento fueron los siguientes:

Instancias correctamente clasificadas	*1015*	*96.6667 %*
Instancias incorrectamente clasificadas	*35*	*3.3333 %*
Error medio absoluto	*0.0081*	
Error medio cuadrático	*0.058*	
Error absoluto relativo	*3.0113 %*	
Error cuadrático relativo	*15.8251 %*	

Al valorar el aprendizaje, los resultados fueron inferiores:

Instancias correctamente clasificadas	*301*	*66.8889 %*
Instancias incorrectamente clasificadas	*149*	*33.1111 %*
Error medio absoluto	*0.13*	
Error medio cuadrático	*0.3132*	
Error absoluto relativo	*47.3463 %*	
Error cuadrático relativo	*83.7776 %*	

Posteriormente fueron modificados los parámetros de entrenamiento y se construyeron diferentes redes, finalmente se encontró una con valores aceptables, que tenía las siguientes características:

- Capas ocultas: 6
- Épocas: 1000
- Tasa de aprendizaje: 0.3
- Momentun: 0.2
- Umbral de validación: 20

Para la fase de entrenamiento se consiguieron los siguientes resultados:

Instancias correctamente clasificadas	*1050*	*100 %*
Instancias incorrectamente clasificadas	*0*	*0 %*
Error medio absoluto	*0.0023*	
Error medio cuadrático	*0.0043*	
Error absoluto relativo	*0.8443 %*	
Error cuadrático relativo	*1.164 %*	

Sin embargo nuevamente se encontraron fallos significantes en la fase de validación:

Instancias correctamente clasificadas	*262*	*58.2222 %*
Instancias incorrectamente clasificadas	*188*	*41.7778 %*
Error medio absoluto	*0.1461*	
Error medio cuadrático	*0.3398*	
Error absoluto relativo	*53.2442 %*	
Error cuadrático relativo	*90.896 %*	

2. *Procesamiento de datos en MatLab*

Ante las altas tasas de error presentadas cuando se usó la herramienta Weka, especialmente en la fase de evaluación, se quiso explorar otras opciones, de esta forma se construyeron diferentes redes neuronales en MatLab usando otras funciones de activación y variando el número de neuronas y de capas ocultas (ver Fig. 5).

Se obtuvieron resultados bastante satisfactorios, a continuación se presentan las dos mejores experiencias:

- Caso 1
  - Capas ocultas: 2 (9 neuronas en cada una)
  - Función de activación: tangente sigmoidal y para la capa de salida una función de activación lineal.
  - Error de entrenamiento: 5e-15
  - Épocas: 2757
  - Gradiente mínimo: 1e-24
  - Tasa de aprendizaje: 0.00001
  - Error de aprendizaje: 0.0071499

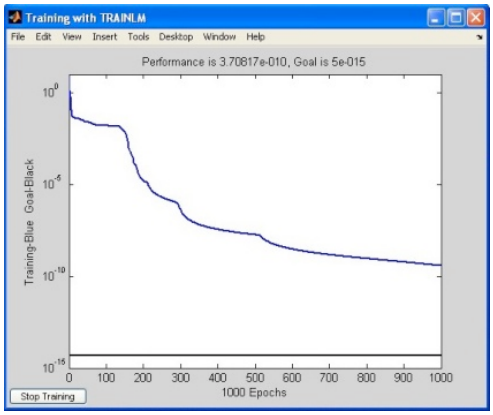

Fig. 5. Gráfica que muestra el desempeño de la red neuronal en Matlab

- Caso 2
  - Capas ocultas: 2 (10 y 9, respectivamente)
  - Función de activación: logaritmo sigmoidal y para la capa de salida una función de activación lineal.
  - Error de entrenamiento: 5e-15
  - Épocas: 1205
  - Gradiente mínimo: 1e-24
  - Tasa de aprendizaje: 0.00001
  - Error de aprendizaje: 2.2504e-014

V. Resultados

Una vez se construyeron los modelos, fue necesario validarlos para evaluar su desempeño en la clasificación de nuevas instancias de datos. Se utilizó la técnica Holdout, el total de datos fue dividido en dos subconjuntos: uno destinado al entrenamiento (70%, es decir 1.050 registros) y el otro se usó para la validación (450 registros). En la tabla 3 se muestran las características de los datos de entrenamiento.

TABLA III
CARACTERÍSTICAS DE LOS DATOS DE ENTRENAMIENTO

Clases	Valores	Cantidad
Erosión hídrica	Muy baja	224
	Baja	206
	Media	206
	Alta	200
	Muy alta	214
Erosión eólica	Muy baja	220
	Baja	210
	Media	213
	Alta	200
	Muy alta	207
Deforestación	Muy baja	209
	Baja	198
	Media	208
	Alta	232
	Muy alta	203
Región climática	Áridas	338
	Semiárida	331
	Subhúmeda seca	381
Desertificación	Muy baja	80
	Baja	285
	Media	500
	Alta	150
	Muy alta	35

VI. Conclusiones

Las herramientas que ofrece Weka presentan ciertas limitaciones en cuanto a opciones para variar los parámetros de los algoritmos, situación que influye sobre los resultados obtenidos.

Las funciones de activación inciden directamente sobre desempeño y los resultados de una red neuronal.

Las técnicas de clasificación y en especial las redes neuronales, son útiles para predecir el grado de desertificación de un suelo a partir de variables incidentes, ya que permiten establecer el grado de influencia de éstas sobre el proceso.

El proceso de minería de datos presenta algunas dificultades, que deben solventarse para obtener buenos resultados: decidir cuál es el algoritmo más apropiado, determinar cuáles modelos deben descartarse porque tienen debilidades y cuáles son óptimos.

Referencias

[1] UNESCO, "Aprendiendo a luchar contra la desertificación". Disponible en: http://unesdoc.unesco.org/images/0012/001258/125816s.pdf

[2] V. Macías, "Hoy es el Día Mundial contra la Desertificación. El hombre irrazonable". Weblog personal de Víctor Macías. Disponible en: http://victormacias.blogia.com/2005/061801-hoy-es-el-dia-mundial-contra-la-desertificacion.php

[3] IDEAM, "Indicadores relacionados con suelos. Tierras afectadas por la desertificación". Disponible en: http://www.ideam.gov.co/indicadores/suelos3.htm

[4] República de Colombia Ministerio de Ambiente, Vivenda y Desarrollo Territorial Viceministerio de Ambiente Dirección de Ecosistemas. "Tercer Informe Nacional de Implementación de la Convención de Naciones Unidas de Lucha Contra la Desertificación". Bogotá, 2006

[5] Mare Terra Fundación mediterránea, "Más de un tercio de la superficie española se encuentra en riesgo de desertificación". Disponible en: http://mareterra.wordpress.com/2007/12/10/

[6] "Los Desiertos ¿Qué son? ¿Cómo se forman?". Disponibleen:http://www.freewebs.com/desertificacion/desiertos.htm

[7] J.E. Rodíguez, "Fundamentos de Minería de Datos". Universidad Distrital Francisco José de Caldas. 2010, pp. 137-138

Sandra Patricia Castillo Landínez: Nació en Bogotá. Es Ingeniera de Sistemas de la Universidad Nacional de Colombia, sede Bogotá. Actualmente adelanta la Maestría en Ciencias de la Información y las Telecomunicaciones con énfasis en Sistemas de Información en la Universidad Distrital Francisco José de Caldas de Bogotá. Se desempeña como docente en la Corporación Universitaria Autónoma del Cauca. Ha trabajado como Asistente de Investigación en varios grupos de la Universidad Militar Nueva Granada, donde participó en diferentes proyectos.
e-mail: sandracastillo@uniatuonoma.edu.co

Pablo Eduardo Caicedo R.: Nacido en Popayán. Ingeniero en Electrónica y Telecomunicaciones de la Universidad del Cauca, candidato a Magister de la misma Universidad. Actualmente se desempeña como docente investigador en la Corporación Universitaria Autónoma del Cauca, coordinador del Grupo de Investigación en Nanoelectrónica y Automatización GRINAU. Su áreas de desempeño son robótica, control e inteligencia artificial.
e-mail: pablocaicedo@uniatuonoma.edu.co

María Isabel Velandia C.: Nació en Bogotá, Colombia. Es Ingeniera Forestal de la Universidad Distrital Francisco José de Caldas, Bogotá, Colombia, donde también obtuvo su título de Especialización en Sistemas de Información Geográfica en el año 2000. En la actualidad adelanta la Maestría en Ciencias de la Información y las Telecomunicaciones con el énfasis en Sistemas de Información en la Universidad Distrital Francisco José de Caldas de Bogotá. Labora como Ingeniera Especialista en Procesamiento Digital de Imágenes de Satélite en el Proyecto SIMCI auspiciado por la Oficina contra las Drogas y el Delito de las Naciones Unidas UNODC y el Gobierno Colombiano a través del Ministerio del Interior y de Justicia. e-mail: isabel.velandia@gmail.com
.