MÉTODOS DE CLASIFICACIÓN EN MINERÍA DE DATOS METEOROLÓGICOS

Methods of Classification in Mining of Meteorological Data

[1]Silvia Haro Rivera*, [1]Lourdes Zúñiga Lema, [2]Antonio Meneses Freire, [1]Luis Vera Rojas, [1]Amalia Escudero Villa

[1]Escuela Superior Politécnica de Chimborazo, Riobamba. Ecuador
[2]Universidad Nacional de Chimborazo, Riobamba, Ecuador

*s_haro@espoch.edu.ec

Resumen

Uno de los objetivos de la minería de datos es la clasificación, la cual tiene como fin clasificar una variable dentro de una de las categorías de una clase. En este trabajo se consideraron variables meteorológicas de la estación Cumandá. Con el objetivo de determinar el modelo adecuado al conjunto de datos, se aplicaron los modelos de clasificación: Naive Bayes, CN2 Rule Induction, K-NN, Tree y Random Forest; así como también los métodos que modifican los parámetros asociados al clasificador: Cross validation, Random sampling, Leave one out y Test on train data. Mediante el software Orange Canvas se calcularon las medidas de rendimiento, Classification Accuary, Precisión Global y Sensibilidad. Se concluyó que los clasificadores Naive Bayes, CN2 Rule Induction y K-NN presentaron valores superiores al 75% de instancias correcta- mente clasificadas. El árbol de decisión y el Bosque Aleatorio superaron el 80%. En cuanto a los métodos que permiten modificar los parámetros asociados al clasificador se determinó que Validación Cruzada, presentó mejores resultados en todas las aplicaciones. La mayor precisión se alcanza en el clasificador bosque aleatorio, con un 83.9% aplicando validación cruzada, seguido por el muestreo aleatorio simple con un porcentaje del 83.1% de verdaderos positivos entre los casos clasificados como positivos.

Palabras claves: métodos de clasificación, minería de datos, datos meteorológicos.

Abstract

One of the colectivamente of data mining is classification, which aims to classify a variable within one of the categories of a class. In this work, meteorological variables of the Cumandá station were considered. In order to determine the appropriate model for the data set, the Naive Bayes, CN2 Rule Induction, K-NN, Tree and Random Forest classification models, as well as Cross validation, Random sampling, Leave one out and Test on train data, that modify the parameters associated with the classifier, were applied. Orange Performance software was used to calculate performance measures, Classification Accuary, Global Accuracy and Sensitivity. It was concluded that the classifiers Naive Bayes, CN2 Rule Induction and K-NN presented values higher than 75% of correctly classified instances. The decision tree and the Random Forest exceeded 80%. Regarding the methods that allow to modify the parameters associated to the classifier, it was determined that Cross-validation presented better results in all the applications. The highest precision is reached in the classifier random forest with 83.9% applying cross-validation, followed by simple random sampling with 83.1% of true positives among the cases classified as positive.

Keywords: classification methods, data mining, meteorological data.

Fecha de recepción: 08-06-2018 Fecha de aceptación: 22-10-2018

I. INTRODUCCIÓN

La minería de datos como herramienta estratégica es clave para explotar el conocimiento de los datos, su objetivo es analizarlos desde todas las perspectivas estratégicas, con el fin de transformar la información y el conocimiento. Mediante la minería de datos se puede: ordenar, clasificar, filtrar y resumir todas las relaciones que un dato puede tener dentro de la información, está centrada no solo en extraer conocimiento sino en encontrar las relaciones o correlaciones que la información; vista desde diferentes (1) ámbitos, tiene con otros datos aparentemente no conectados y, generalmente, recogidos en enormes bases de datos relacionales. La minería de datos en variables meteorológicas tiene una gran aplicación e interés en la actualidad; pues, brinda alternativas diferentes a los métodos tradicionales de análisis y permite estimar variables diversas en casos específicos (2). En este trabajo, primero se realiza un enfoque teórico de cinco clasificadores en minería de datos: Naive Bayes, CN2 Rule Induction, K-NN, Tree y Random Forest; así como, los parámetros para evaluar el rendimiento de cada uno de estos y los métodos que permiten modificar al clasificador. Segundo, mediante el *software* Orange Canvas se realiza la aplicación; y tercero, se analizan los resultados y emiten conclusiones.

Clasificadores

La clasificación en minería de datos es una técnica supervisada, donde generalmente se tiene un atributo llamado clase y se busca determinar si los atributos pertenecen o no a un determinado concepto (3).
La clasificación, es la habilidad para adquirir una función que mapee (clasifique) un elemento de dato a una de entre varias clases predefinidas. Un objeto se describe a través de un conjunto de características (variables o atributos) $X \rightarrow \{X_1, X_2, \ldots, X_n\}$. El objetivo de la tarea de clasificación es clasificar el objeto dentro de una de las categorías de la clase $C = \{C_1, \ldots, C_k\}$

$$f: X_1 \times X_2 \times \ldots \times X_n \rightarrow C$$

Las características o variables elegidas dependen del problema de clasificación. Para el estudio se consideraron los siguientes clasificadores:

Naive Bayes

La clasificación mediante el algoritmo Bayesiano ofrece la solución óptima de la probabilidad de pertenencia de cada muestra a todas las clases. De acuerdo a la teoría general de la probabilidad de Bayes, dado x el objetivo es asignar x a alguna de las clases existentes. Supóngase que las clases son $S_1, S_2, \ldots, S_c$; para cada una de ellas se puede estimar la función de densidad de probabilidad, por lo que para la observación x se trata de determinar la probabilidad *a posteriori* de que dicha muestra pertenezca a la clase S_j, misma que se puede calcular por:
Donde (m_j, C_j) es la función de densidad

$$P(S_j \mid x) = \frac{P(x \mid m_j, C_j)P(S_j)}{\sum_{(j=1)}^{c} P(x \mid m_j, C_j)}$$

de probabilidad para la clase S_j (4).

CN2 Rule Induction

El algoritmo CN2 es una técnica de clasificación diseñada para la inducción eficiente de reglas sencillas y comprensibles de forma "si entonces", este modelo funciona solo para clasificar. La búsqueda de reglas puede ser por:

- Medida de evaluación, selecciona una regla heurística para evaluar las hipótesis encontradas; puede darse por: entropía (medida de la imprevisibilidad del contenido), mediante la precisión de Laplace o por una precisión relativa ponderada.
- Ancho del haz, recuerda la mejor regla encontrada hasta el momento y monitorea un número fijo de alternativas.

K-NN o Nearest Neighbours (Vecinos más cercanos)

El método K-NN emplea la clasificación supervisada estimando la distancia de cierto número de muestras (K vecinos)

a la muestra que se pretende clasificar, determinando su pertinencia a la clase de la que encuentre más vecinos etiquetados, considerando el criterio de mínima distancia. Esta técnica es válida solo para datos numéricos, no para clasificadores de textos (5). Dado un conjunto de muestras $X = \{x_1, x_2, \ldots, x_n\} \epsilon R^p$ con función de distancia d, se tiene:

1. Vecino más cercano: localizar la muestra x_l en X más cercana a x_k, con $l \leq k \leq N$ y $k \neq 1$.
2. Rango r: dado un umbral r y n punto x_k, no considerará los puntos x_l que satisfacen $0 \leq d(x_k, x_l) = d_{kl} \leq r$.

Tree (Árboles de decisión)

Es una técnica de clasificación supervisada (6), permite determinar la decisión que se debe tomar siguiendo las condiciones que se cumplen desde la raíz hasta alguna de sus hojas (7). El árbol de decisión se construye partiendo el conjunto de datos en dos o más subconjuntos de observaciones, después estos subconjuntos se vuelven a particionar empleando el mismo algoritmo. La raíz del árbol es el conjunto de datos inicial, los subconjuntos y subsubconjuntos conforman las ramas del árbol. El conjunto en el que se realiza una partición se llama nodo y permite bifurcar en función de los atributos y sus valores. Las hojas del árbol proporcionan predicciones. Los algoritmos más utilizados son: ID3, C4.5 y CART (8)

Random Forest (Bosque aleatorio)

Es una combinación de árboles predictivos, el cual trabaja con una colección de árboles incorrelacionados y los promedia; de modo que cada árbol depende de los valores de un vector aleatorio de la muestra de manera independiente y con la misma distribución de todos los árboles en el bosque (6). Random Forest o "Selvas Aleatorias" es una técnica predictiva en la cual todos los clasificadores del método del consenso (Bagging) son árboles de decisión. Cada modelo genera una predicción y se selecciona por la mayor cantidad de votos (8).

Evaluación de los clasificadores

Para evaluar los clasificadores se consideraron los siguientes parámetros (9):

Classification acccuary (CA): Determina la proporción de ejemplos correctamente clasificados.

Precisión y exactitud: La primera se refiere a la dispersión del conjunto de valores obtenidos de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión; y la segunda, se refiere a cuán cerca del valor real se encuentra el valor medido. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación. Cuanto menor es el sesgo más exacta es una estimación. Estos indicadores de precisión se pueden calcular por:

$$Precisión = \frac{tp}{tp + fp}$$

$$Exactitud = \frac{tp + tn}{tp + tn + fn + fp}$$

Recall (sensibilidad o exhaustividad): es la proporción de verdaderos positivos entre todos los casos positivos en los casos. Si su resultado es uno entonces se han encontrado verdaderos positivos en la base de datos, por lo que no existiría ruido ni silencio informativo. Por el contrario si su valor es cero los datos no poseen relevancia alguna. Se calcula por:

$$Recall = \frac{tp}{tp + fn}$$

Matriz de confusión

La matriz de confusión permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Las columnas representan el número de predicciones de cada clase, mientras que cada fila representa las instancias en la clase real (9). A continuación se muestra la matriz de confusión para el caso donde se tienen dos clases.

		Predicción	
		Negativo	Positivo
Valor Real	Negativo	VN (verdadero negativo)	FP (falso positivo)
	Positivo	FN (falso negativo	VP (verdadero positivo

Tabla 1. Matriz de confusión, dos clases

Métodos para modificar los parámetros

Los métodos utilizados para modificar los parámetros asociados al clasificador fueron:

Cross validation: validación cruzada es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba (10). En Orange Canvas el algoritmo divide al conjunto de datos en pliegues (generalmente entre 5 o 10). El algoritmo se prueba manteniendo ejemplos de un pliegue a la vez, el modelo se induce de otros pliegues y se clasifican ejemplos del pliegue retenido. Esto se repite para todos los pliegues.

Random sampling (muestreo aleatorio simple): divide aleatoriamente los datos en el entrenamiento y el conjunto de pruebas en la proporción fijada por el usuario, se repite el proceso durante el número especificado de veces.

Leave one out: es similar al cross validation pero tiene una instancia a la vez, induce el modelo de todos los demás y luego clasifica las instancias presentadas. Este método es muy estable, fiable pero lento.

VARIABLE	DESCRIPCIÓN
DÍA	Día
MES	Mes
AÑO	Año
HORA	Hora
TEMP	Temperatura
HUMREL	Humedad relativa
PREBARO	Presión Barométrica
RADIFU	Radiación difusa
RSG	Rediación solar global
TEPSU	Temperatura del suelo
VV	Velocidad de viento

Tabla 2. Variables de estudio

Test on train data: utiliza todo el conjunto de datos para el entrenamiento y luego para la prueba; este método siempre da resultados erróneos.

Test on test data: este test permite introducir otro conjunto de datos con ejemplos de prueba (desde otro archivo o seleccionados en otro enlace).

II. METODOLOGÍA

Los datos empleados en este trabajo corresponden a registros por día y hora (00:00 a 23:00) durante los 12 meses del año 2015. El fichero de datos contiene 92 136 observaciones y las variables se detallan en la tabla 2.

Se generó la variable categórica ESCALA considerando intervalos de tiempo con las siguientes restricciones:

- Escala-I: horas comprendidas entre la 01:00 y 06:00
- Escala-II: horas comprendidas entre la 07:00 y 12:00
- Escala-III: horas comprendidas entre las 13:00 y 18:00
- Escala-IV: horas comprendidas entre las 19:00 y 00:00

El objetivo del análisis es aplicar las técnicas de clasificación en la base de datos descrita, mediante el programa Orange Canvas y determinar los factores que caracterizan los grupos horarios propuestos. Los pasos aplicados fueron los siguientes:

1. Carga del fichero en el *software* Orange Canvas
2. Selección de la variable objetivo de estudio ESCALA
3. Aplicación de los métodos de clasificación
4. Evaluación de los métodos de clasificación

Clasificador	Test de prueba	C.A. (%)	Precisión	Recall (5)
Naive Bayes	Cross validation	78,0	74,3	72,2
	Random samping	77,8	75,0	77,8
	Leave one out	77,9	75,2	72,1
	Test on train data	75,3	77,9	72,3
CN2 Rule Induccion	Leave one out	78,2	73,2	80,2
	Test on train data	100	100	100
K-NN	Cross validation	78,6	81,0	79,7
	Random samping	77,9	80,6	79,6
	Leave one out	78,5	80,8	80,0
	Test on train data	100	100	100
Tree	Cross validation	82,1	80,3	78,4
	Random samping	81,7	80,0	78,8
	Leave one out	82,3	80,2	78,0
	Test on train data	96,2	96,4	94,7
Random Forest	Cross validation	84,7	83,9	81,1
	Random samping	84,7	83,1	81,9
	Leave one out	84,9	83,9	82,4
	Test on train data	97,6	97,4	96,5

Tabla 3. Resultados de los clasificadores.

III. RESULTADOS Y DISCUSIÓN

Los resultados obtenidos se muestran en la tabla 3; donde se pueden observar los porcentajes obtenidos en los tres parámetros de evaluación de cada uno de los clasificadores empleados y para cada prueba.

En la tabla 4, se muestra las matrices de confusión de los clasificadores: Naive Bayes (NB), K-NN, Random Forest (RF) y Tree, (T); para el caso particular donde el parámetro que modificó al clasificador fue Cross validation.

La matriz de confusión (Tabla 4) muestra que, para el modelo Naive Bayes de los 2094 datos del grupo escala-I; 1678 se clasificaron correctamente, 398 se han clasificado en escala-II y 18 en escala-IV. Además; 1511 de escala-II, 1726 de escala-III y 1617 de escala IV también se han clasificado correctamente dentro de sus respectivos grupos con igual número de datos. En el clasificador K-NN, se puede observar que el mayor porcentaje de datos correctamente agrupados en su grupo corresponde a escala-I con un valor del 84.1% y el menor es de escala IV con el 68.7%. En Random Forest (RF), el mayor número de datos correctamente clasificado en su grupo es escala-III; y en el modelo Tree (T), fue escala-III, seguido por escala-I.

El árbol de decisión (tree) se generó con cuatro niveles de profundidad y se muestra en la figura 1.

		PREDICCIÓN															
		escala-I				escala-II				escala-III				escala-IV			
		NB	K-NN	RF	T	NB	K-NN	RF	T	NB	K-NN	RF	T	NB	K-NN	RF	T
Valor actual	escala-1	1678	1761	1813	1744	398	332	276	339	0	0	0	0	18	1	5	11
	escala-II	443	379	356	410	1511	1669	1669	1642	132	18	21	19	8	28	18	23
	escala-III	0	0	0	0	41	13	14	21	1776	1718	1844	1772	327	363	236	301
	escala-IV	124	75	9	16	57	46	35	42	296	535	311	315	1617	1438	1739	1721

Tabla 4. Matriz de confusión: Naive Bayes, K-NN, Random Forest y Tree. Cross validation

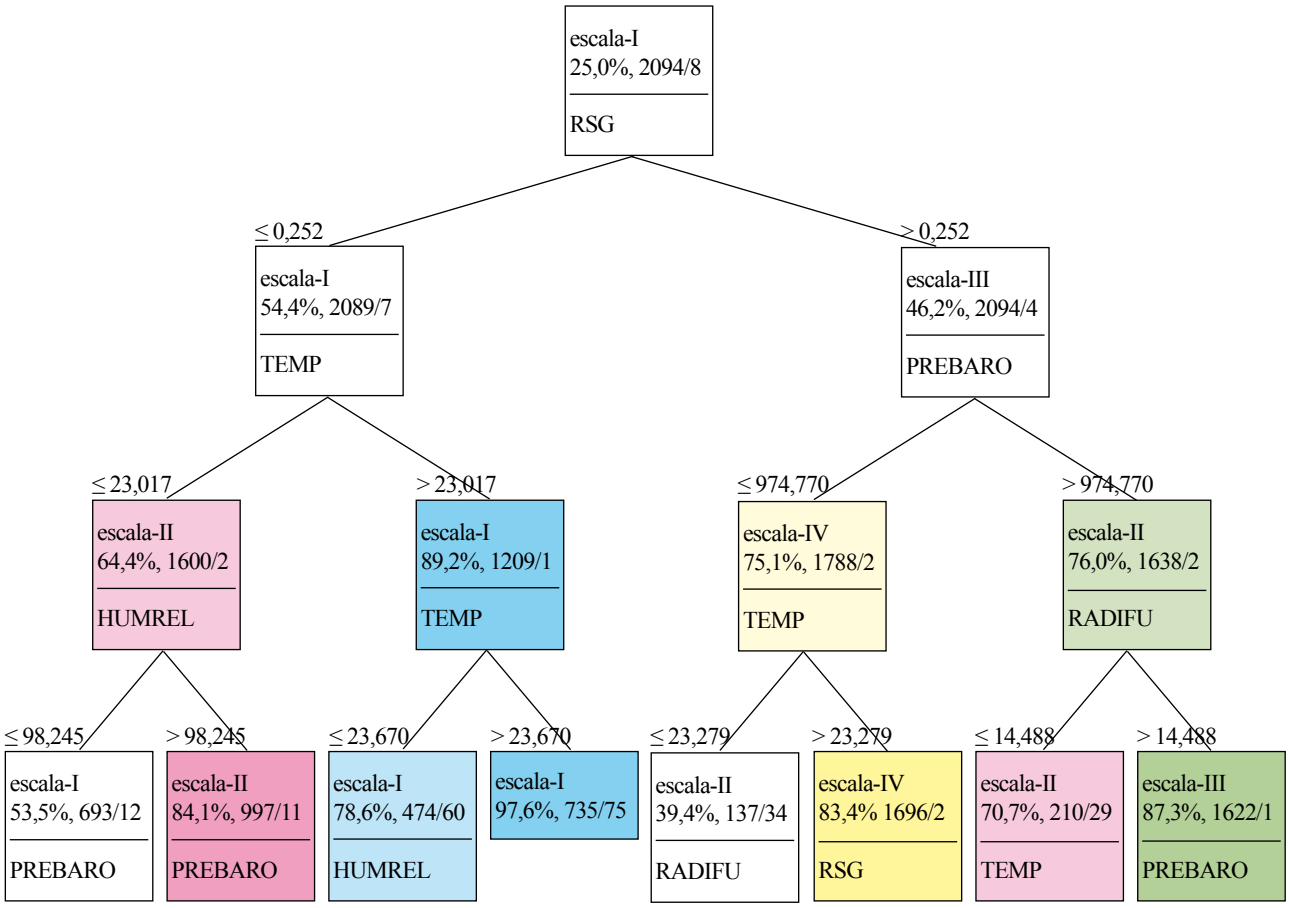

Figura 1: Árbol de decisión

De acuerdo al árbol obtenido se puedo establecer que: si la radiación solar global (RSG) es mayor que 0.252 entonces predice la presión barométrica (PRESIBARO) dentro de la escala III; esto es, en el horario de 13:00 a 18:00. De similar forma, si la radiación solar global es menor o igual que 0.252 predice la temperatura (TEMP), en el horario de 01:00 a 06:00 (escala-I). De igual manera se puede interpretar las ramas inferiores del árbol.

IV. CONCLUSIONES

De la tabla 2 podemos concluir que los clasificadores Naive Bayes, CN2 Rule Induction y K-NN presentan valores superiores al 75% de instancias correctamente clasificadas. El árbol de decisión (Tree) y el bosque aleatorio (Random forest) superaron el 80%. En cuanto a los métodos que permiten modificar los parámetros asociados al clasificador se pudo determinar que validación cruzada (Cross validation), es el que presenta mejores resultados en todas las aplicaciones. La mayor precisión se presentó en el bosque aleatorio. Los resultados del test on train data en los clasificadores CN2 Rule Induction y K-NN son del 100%, pero no son confiables pues emplean todo el conjunto de datos para el entrenamiento y luego para la prueba. La mayor precisión se alcanza en el clasificador bosque aleatorio, con un 83.9% aplicando validación cruzada, seguido por el muestreo aleatorio simple con un porcentaje del 83.1% de verdaderos positivos entre los casos clasificados como positivos. En el mismo caso se determinó que la proporción de verdaderos positivos entre todos los casos positivos en los casos (sensibilidad) es mayor en el mismo clasificador. Se pudo evidenciar que los resultados en los clasificadores mediante los diferentes métodos no varían en proporciones significativas.

V. AGRADECIMIENTO

Al Centro de Investigaciones de Energía Alternativa y Ambiente de la Espoch, Facultad de Ciencias.

Referencias

1. Hernández E, Duque N, Cadavid J. Big Data: una exploración de investigaciones, tecnologías y casos de aplicación. TecnoLógicas. 2017.
2. Duque N, Orozco M. Minería de Datos para el Análisis de Datos Meteorológicos. Tendencias en Ingeniería de Software e Inteligencia Artificial. ;: p. 105-114.
3. Segrera S, Moreno M, Miguel L. Aplicación de la minería de datos en la evaluación de la aptitud física de las tierras para el cultivo de la caña de azúcar. III Taller Nacional de Minería de Datos y Aprendizaje. 2005;: p. 349-358.
4. Sandoval Z, Prieto F. Caracterización de café cereza empleando técnicas de visión artificial. Facultad Nacional de Agronomía-Medellín. 2007;: p. 4105-4127.
5. Pascual D, Pla F, Sánchez S. Algoritmos de Agrupameinto. Revista Facultad de Ingeniería. 2008;: p. 163-175.
6. Medina R, Ñique C. Bosques Aleatorios como extensión de los árboles de clasificación con los programas R y Python. Interfases. 2017;: p. 165-189.
7. Robles Y, Sotolongo A. Integración de los algoritmos de minería de datos 1R, PRISM E ID3 A POSTGRESQL. Gestión de Tecnología y Sistemas de Información. 2013;: p. 389-406.
8. Ochoa L, Paredes K, Araya C. Evaluación de Técnicas de Minería de Datos para la Predicción del Rendimiento Académico. Global Partnerships for Development and Engineering Education. 2017.
9. Graham W. Data Mining with Rattle and R New York, USA: Springer; 2011.
10. Orozco E, García DA. Métodos de clasificación para identificar lesiones en piel a partir de espectros de reflexión difusa. Revista Ingeniería Biomédica. 2010;: p. 34-40.