Proyecto Predicción de Estado Civil

Nobrega Yanelly y Prieto Wilmer

Escuela de Computación
Facultad de Ciencias
Universidad Central de Venezuela (UCV)
Av. Paseo Los Ilustres, Los Chaguaramos, Caracas, Venezuela
yanellynobrega@gmail.com, wilmerprieto3000@gmail.com

Resumen

Este proyecto está basado en una competencia de Kaggle [1], que tiene como objetivo la búsqueda de un algoritmo que permita generar un modelo que prediga de manera eficiente el estado civil de una pareja dado un set de datos el cual contiene una serie de características de la pareja, tales como: la edad del hombre, edad de la mujer, ingreso anual de ambos, entre otros. Todos estos datos pasarán por un pre-procesamiento y finalmente para la generación de los modelos se utilizarán los algoritmos Árbol de Decisión, K-Vecinos y Reglas de Clasificación.

Palabras Clave: predicción, pre-procesamiento, clasificación, estado civil, pareja, árbol de decisión, K-Vecinos.

1. Introducción

Existen diversos algoritmos de clasificación basados en un conjunto de cálculos y reglas que permiten crear un modelo a partir de los datos. Para la creación de estos modelos, el algoritmo analiza los datos proporcionados en busca de tipos específicos de patrones o tendencias y usa los resultados de este análisis para definir los parámetros óptimos para la creación del modelo. A continuación, estos parámetros se aplican en un conjunto de datos de entrenamiento para extraer patrones.
Para este proyecto utilizaremos 3 algoritmos de clasificación (Árbol de Decisión, K-Vecinos y Reglas de Clasificación) con el fin de predecir eficientemente el estado civil de una pareja dado un set de datos con características que determinan si la pareja esta separada, casada o divorciada.

2. Explicación del Problema

Se plantea el objetivo de predecir el estado civil de una pareja dado un set de datos. Dicho set de datos posee siete columnas que son: **ID** - Identificador unico para cada registro(pareja), **GAGE** - Edad de la Mujer, **BAJE** - Edad del Hombre, **GP**, **BP**, **AINCOME** - Salario Anual en dolares percibido por la pareja y **STATUS** - Estado civil de la pareja(Atributo a predecir). Ya identificadas las columnas se procederá con el siguiente paso que es el pre-procesamiento de los datos.

3. Pre-procesamiento de Datos

Para realizar el pre-procesamiento de la data se realizó un análisis exploratorio de los mismos en donde se observó que existían columnas irrelevantes en cuanto al objetivo del problema, dichas columnas fueron el ID y AINCOME los cuales representan el identificador de la fila y el sueldo anual de la pareja respectivamente. En cuanto a este último atributo lo consideramos irrelevante ya que no hallamos ningún tipo de patrón que nos encaminara a predecir el valor real del estatus de la pareja. Seguido de esto se decidió aplicar una numeración a las columnas que contenían variables nominales las cuales podían asumir n valores, dichas columnas son: GP, BP y STATUS. Cabe destacar que al realizar el análisis se pudo concluir que la diferencia entre ambas edades era un factor fuerte para determinar el estado civil debido a que se observaron diversas similitudes de diferencia de edad para un mismo valor de clase, por lo que se procedió a eliminar ambas columnas y crear una nueva almacenando la diferencia de edades.

4. Modelos de Clasificación

Los algoritmos utilizados para lograr predecir el estado civil de una pareja con los datos obtenidos y ya pre-procesados fueron Árbol de Decisión, K-Vecinos y Reglas de Clasificación. Se decidió utilizar estos métodos de clasificación porque fueron los estudiados en la materia de Minería de Datos y de los cuales se tenía mayor conocimiento a la hora de implementar la solución al problema planteado. Además es bueno destacar que son métodos con un funcionamiento fácil de entender, y dado que cada algoritmo de estos se basa en diferentes principios, permite que dependiendo del problema algunos modelos sean mejores que otros, ampliando así la posibilidad de lograr un buen resultado. Para la evaluar la eficiencia de los modelos se tomó en cuenta la matriz de confusión y la curva ROC.
La matriz de confusión[2] es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real.
Las curvas ROC (Receiver Operating Characteristic)[3] representan la sensibilidad de una prueba diagnóstica que produce resultados continuos, en función de los falsos positivos (complementario de la especificidad), para distintos puntos de corte.

4.1 Árbol de Decisión

Un árbol de decisión[4] es un modelo de predicción en el cual dada una base de datos se generan diagramas de construcciones lógicas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.
Se generó el modelo para el árbol de decisión usando la función rpart, cuyos parámetros fueron: (STATUS .) el cual denota que se quiere predecir la característica STATUS (estado civil) de la pareja basándose en las demás características provistas en el set de datos y el parámetro method: (class) indicando se quiere un modelo de clasificación. En cuanto a los parámetros de control: minsplit, cp, maxdepth (los cuales nos indican la cantidad mínima de individuos por nodo, el grado de brusquedad en los cambios y la profundidad de árbol respectivamente) no fueron utilizados debido a que en al realizar la permutación de los mismos, los resultados del árbol fueron muy similares. Para la evaluación de este modelo se tomó en cuenta la matriz de confusión y el área bajo la curva (función roc del paquete pROC) obteniendo como resultado un noventa (90) por ciento de aciertos y 0.97 de área bajo la curva respectivamente.

4.2 K-Vecinos

El método k-vecinos mas cercanos[5] es un método de clasificación supervisada que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento pertenezca a la clase a partir de la información proporcionada por el conjunto de prototipos.
Se generó el modelo de K-Vecinos haciendo uso de la función knn, pasando los conjuntos de prueba y entrenamiento, y el parámetro k (número de vecinos considerados) asignándole el valor 28 debido a que se tomó como medida la raíz cuadrada del número total de individuos en el set de datos (750) que suele ser la medida que mejor se comporta teóricamente. Para la evaluación de este modelo se tomó en cuenta la matriz de confusión y el área bajo la curva (función roc del paquete pROC) obteniendo como resultado un cincuenta (50) por ciento de aciertos y 0.62 de área bajo la curva respectivamente.

4.3 Reglas de Clasificación

Una regla de clasificación[6] es un procedimiento en el que cada elemento de una población es asignado a una de las clases, donde la población es un conjunto de miembros que pueden ser agrupados en n conjuntos denominados clases.
Se generó el modelo de Reglas de Clasificación haciendo uso de la función JRip provista por la interfaz RWeka, para ello debimos realizar una transformación de los datos haciendo uso de **as.factor**, y luego aplicamos dicha función JRip, donde el primer parámetro es formula: (STATUS .) el cual denota que se quiere predecir la característica STATUS de la pareja basándose en las demás características provistas en el set de datos, y el segundo parámetro que indica la data a utilizar que en este caso es la data de entrenamiento. Para la evaluación de este modelo se tomó en cuenta la matriz de confusión y el área bajo la curva (función roc del paquete pROC) obteniendo como resultado un cien (100) por ciento de aciertos y 1 de área bajo la curva respectivamente.

5. Conclusión

Como mencionamos anteriormente para realizar la comparación de los resultados obtenidos entre los tres modelos implementados, se tomó en cuenta la matriz de confusión para obtener la tasa de aciertos y el área bajo la curva (función roc del paquete pROC) de cada uno de ellos. Se tiene que el mejor modelo obtenido es el basado en Reglas de Clasificación debido a que la tasa de aciertos es perfecta al igual que el área bajo la curva, es decir, para ambos casos el resultado fue de 1. Seguidamente el modelo basado en árbol de decisión con una tasa de aciertos de 0.91 y con un área bajo la curva de 0.97. Y por último se tiene el modelo de K-Vecinos el cual su tasa de aciertos fue de 0.5 y el área bajo la curva de 0.62 por lo que queda totalmente descartado como mejor modelo. Sin duda alguna el mejor modelo para el caso estudiado es el modelo basado en Reglas de Clasificación.

6. Agradecimientos

Agradecemos al Profesor Fernando Crema, quien se encargo de instruirnos en el área de Minería de Datos específicamente en lo referente a reglas de clasificación lo cual nos ayudó a elaborar dicho estudio y también al preparador de Minería de Datos Wilmer Gonzalez quien estuvo atento a cualquier duda que se nos presto para ayudarnos.

Referencias

[1] Marriage Prediction, Kaggle Competitions. Disponible en: https://inclass.kaggle.com/c/marriage-prediction/

[2] Confusion Matrix. Disponible en: http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html

[3] ROC Graph. Disponible en: http://www2.cs.uregina.ca/~dbd/cs831/notes/ROC/ROC.html

[4] Lior Rokach and Oded Maimon (2008). Data mining with decision trees: theory and applications. World Scientific.

[5] Fix, E.; J.L. Hodges (1989). «(1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951)»

[6] MathWorld article for statical test. Disponible en: http://mathworld.wolfram.com/StatisticalTest.html