Uso de R y minería de datos para la predicción del estado de cultivos cítricos

Martín Ehman[1], Gabriel Surraco[1], Karina Eckert[1], Sergio Garrán[2], Vanesa Hochmaier[2], y Armando Taie[3]

[1] Universidad Gastón Dachary. Posadas, Misiones
{martinehman90, gabrielsurraco1, karinaeck}@gmail.com
[2] INTA EEA Concordia. Concordia, Entre Ríos
sergiomariogarran@gmail.com hochmaier.vanesa@inta.gob.ar
[3] INTA EEA Corrientes. El Sombrerito, Corrientes
taie.armando@gmail.com

Palabras Clave: Lenguaje R · Minería de Datos · Predicción de estado de cultivos · Modelos de predicción de rendimiento · FruTIC.

1. Introducción

El estado de sanidad de los cultivos es un factor primordial en la comercialización de cítricos en la región y el ámbito internacional, más aún en este último donde la barrera de entrada es más alta. Por ello es importante implementar estrategias que aseguren calidades óptimas de los cítricos aplicando estrategias de manejo integrado de cultivos.
FruTIC [**1**] es una herramienta inteligente que permite determinar e informar a los productores y técnicos los momentos óptimos para las labores en los cultivos, permitiendo así una utilización eficiente de recursos. Este sistema se basa en el concepto del triángulo de la enfermedad y reúne y provee información sobre las variables meteorológicas, poblaciones de plagas y enfermedades, y etapas de desarrollo de los cultivos. La misma contiene información histórica que ha sido captada a partir del año 2009. La predicción del estado de los cultivos cítricos se realizó mediante técnicas de Minería de Datos (MD) implementadas en lenguaje R bajo el entorno *RStudio* [**2**], con el fin de permitir el conocimiento anticipado del estado de las plantas y posibilitar la realización de acciones puntuales sobre los lotes para minimizar las consecuencias negativas que puedan alcanzar al producto final.

2. Materiales y métodos

Utilizando la metodología CRISP-DM [**3**], en la fase del entendimiento de datos se obtuvo un panorama general de las estructuras de datos y los atributos de los *datasets*. Algunos de los paquetes utilizados en esta fase fueron *ggplot2* para visualizaciones y *data.table* para lectura rápida de archivos. Una

vez que se adquirió un primer conocimiento de los datos se procedió al preprocesamiento de datos. Entre las tareas específicas se realizaron transformaciones de tipos de datos, reestructuración de *dataframes*, exploración y corrección de valores anómalos, imputación de datos faltantes, integración de datos, selección de variables y finalmente un análisis del *dataset* consolidado incluyendo análisis multivariados y de correlaciones. Entre los paquetes utilizados en esta fase se pueden mencionar *qdap*, *dplyr*, *plyr*, *survival*, *stringr*, *lubridate* para el manejo de fechas, *dummies*, *gridExtra* y *ggplot2* para visualizaciones, *mice* e *imputeTS* para imputación de datos y *naniar* para la exploración de datos faltantes.
En las fases de modelado y evaluación el principal paquete utilizado fue *caret*. Con *caret* se realizó la validación cruzada, el entrenamiento y la evaluación de modelos. Otros paquetes utilizados en estas fases fueron *caretEnsemble*, *mlbench* para *benchmarks*, *gmodels* para matrices de confusión y *pROC* para la obtención del área bajo la curva ROC.
Se puede observar en la Tabla 1 los resultados obtenidos en los modelos de predicción. Los modelos que mejor desempeño obtuvieron en términos de área bajo la curva ROC fueron *gbm* y *xgboost*.

Tabla 1. Resultados sobre los datos de prueba

Modelo	Precisión	Kappa	Área bajo la curva ROC
C5.0	0,9164	0,7221	0,6797
gbm	0,8885	0,5849	0,7582
knn	0,9019	0,6771	0,6289
nnet	0,8130	0	0,6055
rf	0,9137	0,7269	0,7441
rpart	0,8272	0,1314	0,6683
xgbTree	0,9153	0,7188	0,7910

Por otro lado se obtuvo la importancia de los atributos predictores para el modelo seleccionado (*xgboost*). Se han detectado varios atributos relevantes al modelo, entre ellos: humedad, temperatura y velocidad del viento, estadios de brotación y lotes particulares. La obtención de dichas importancias es relevante para tener un conocimiento más detallado de las relaciones entre las variables del triángulo de las enfermedades y el estado de los cultivos.

Referencias

1. Stablum A. et al.: FruTIC: Sistema interactivo que permite un manejo integrado del cultivo cítrico. 2º Congreso de Agroinformática - 39º Jornadas Argentinas de Informática, pp. 680-695, Buenos Aires (2010)
2. RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA, http://www.rstudio.com
3. Chapman P. et al.: CRISP-DM 1.0: Step by step data-mining guide, https://www.the-modeling-agency.com/crisp-dm.pdf