APLICACIÓN DE MINERÍA DE DATOS PARA DETECCIÓN DE PATRONES EN INVESTIGACIONES BIOTECNOLÓGICAS

Ruiz Omar[1], Bauz Sergio[2], Jiménez María[3]
[1 2 3] Escuela Superior Politécnica del Litoral
Campus Galindo, Km. 30.5 vía Perimetral
Apartado 09-01-5863, Guayaquil, Ecuador
Email: oruiz@espol.edu.ec

RESUMEN

El presente trabajo despeja la interrogante de la aplicabilidad de la minería de datos, en estudios de diferentes variedades de banano ubicadas en varias zonas geográficas. Aborda cada una de las etapas que conlleva la preparación de los datos, la creación del repositorio de la data, y una aplicación amigable que hace uso de Excel y sus tablas dinámicas cuando se realizan las consultas; para los algoritmos de minería de datos, se utilizaron librerías del software estadístico R. Se determinó que los datos deben ser muy trabajados para poder obtener información de ellos; la minería de datos es aplicable para este tipo de estudios; finalmente se detectaron patrones biológicos estables en las variedades estudiadas.

INTRODUCCIÓN

La poca disponibilidad de tiempo para obtener información con sustento estadístico, crea la necesidad de aplicar nuevas metodologías de análisis para estudios biológicos realizados especialmente en campo.

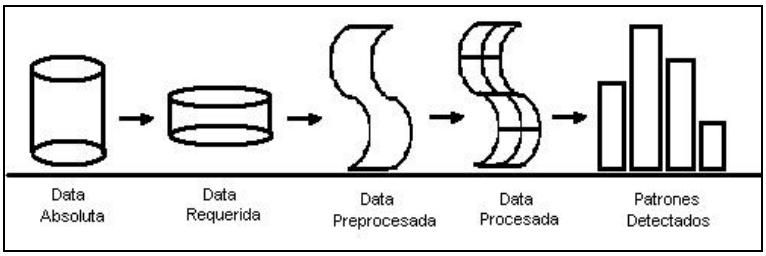
Cuando los datos son analizados de forma tradicional, la información a conseguir es limitada. Análisis estadísticos matemáticos complejos estarán ausentes; lo que genera escasa valoración, desperdicio de información importante, que por falta de una herramienta informática más eficiente y amigable, no se pueda explorar a fondo para encontrar patrones conductuales de los entes investigados. Es claro que esta situación limita al investigador al momento de tomar decisiones al no contar con información ágil, fidedigna, demostrable y estadísticamente sustentada.

Para realizar lo anteriormente expuesto se aplica Minería de Datos (MD); el desarrollo de una aplicación informática que utilice librerías del software R, ayudará a detectar patrones biológicos.

Se utilizarán datos obtenidos en estudios realizados por el Centro de Investigaciones Biotecnológicas del Ecuador (CIBE-ESPOL), sobre variables agronómicas, fitosanitarias y geográficas, obtenidas desde el 2004 hasta el 2006, de plantaciones bananeras de este sector productivo de la costa ecuatoriana.

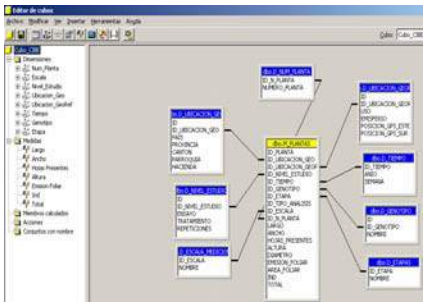
MATERIALES Y MÉTODOS

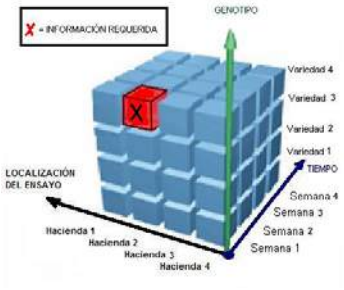
La primera etapa del proyecto consistió en construir la base de datos (BD) y almacenar la data para obtener su integridad, validez, relevancia y confiabilidad, a través de la normalización de los mismos y su posterior pre-procesamiento, siguiéndose el esquema conceptual planteado a continuación:



Se investigaron diferentes herramientas informáticas para desarrollar el Data Warehouse (DW), se seleccionó MS SQL-SERVER. Además se decidió hacer uso de las herramientas de Microsoft tales como: Visual Basic y Servicios OLAP entre otros.

Creada la BD multidimensional, se procedió a elaborar el diseño físico del Cubo de Datos (DC), mediante un esquema en estrella y con dimensiones variables de procesamiento. Luego, para dar facilidades de manipulación supervisada de los datos, hacia los usuarios de la aplicación, se crea un enlace entre el Analisys Services y la herramienta de escritorio, Microsoft Excel, de tal manera que el DC pueda ser administrado desde una tabla dinámica.





Se realizó la recopilación de la data histórica para ser depositadas en la nueva BD, se diseñó y desarrolló un Sistema para Transferencia de Datos (DTS) considerando las validaciones en cuanto a los comportamientos estadísticos de las variables, valores máximos y mínimos permisibles, si son crecientes en el tiempo o son series temporales, etc. bajo estas restricciones, se procedió a la transferencia de los datos.

Se realizó la depuración y validación, luego se procedió con el Preprocesamiento o Preparación de los datos. Este paso es muy importante, porque en él se debe decidir que hacer con los valores perdidos (*missing values*), cual será la regla por la cual se haga inferencia de dichos valores si se decide completar la data faltante; para ello existen diferentes metodologías en dependencia del tipo de datos y su comportamiento estadístico; por ejemplo, si la variable es creciente en el tiempo, se puede hacer interpolación cruzada o regresión lineal; si es una variable que fluctúa en el tiempo, se puede utilizar modelos de series temporales; si sus valores giran alrededor de un valor central, se puede utilizar la media de los valores conocidos, etc.

Se detectó la necesidad de crear nuevas variables, a través de funciones de variables aleatorias, las cuales ofrecen mayor información, tal es el caso de la aplicación del **área bajo la curva** para analizar la evolución en el tiempo de variables que describen características especiales como la altura o el diámetro de una planta, desde la siembra hasta la cosecha.

Finalizada la construcción de la BD, la aplicación del DC y del DTS, y una vez normalizados y preprocesados los datos, se procede a la selección de las metodologías estadísticas básicas y la selección de los algoritmos de MD.

Para la realización del análisis exploratorio de los datos, son indispensables las técnicas estadísticas clásicas de para el análisis descriptivo, además estadística inferencial a través de contrastes de hipótesis. La siguiente etapa es la Estadística predictiva a través de la obtención de modelos matemáticos que describan el comportamiento de los datos, para ello se puede aplicar análisis de regresión o multivariado.

Seguidamente se deben seleccionar los métodos o técnicas de clasificación para la MD en base a la aplicación de métodos estadísticos robustos y comprobados.

Ya que la idea principal del presente estudio es proveer información ágil con soporte estadístico pero de fácil interpretación para los investigadores, se consideraron las metodologías de menor complejidad al momento de analizar los resultados.

Como técnica de agrupamiento se utilizo el clustering. Una buena alternativa es el clustering No supervisado – jerárquico, porque ofrece la posibilidad de mostrar a través de Dendogramas el agrupamiento de los casos o registros, no así el clustering No supervisado – No jerárquico, pues este no da esa posibilidad; y La Correlación, como método para realizar agrupamiento por afinidad.

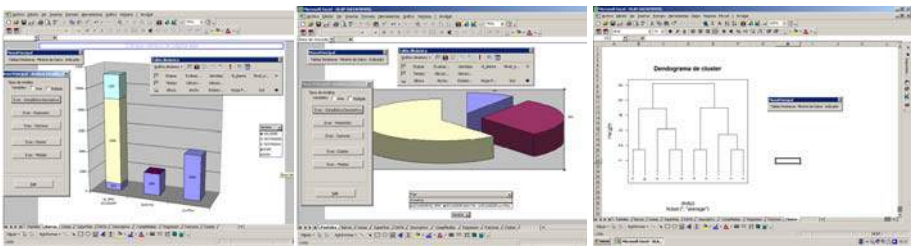
El análisis de regresión fue considerado para realizar predicciones. El análisis discriminante como método de clasificación, pues este asigna un caso o registro a uno de los diferentes grupos previamente definidos en base a información histórica.

Otra técnica estudiada es la “Regla de inducción”, (extracción de reglas if-then de datos basados en significado estadístico) identifica elementos de las poblaciones estudiadas que pudiesen responder de manera similar ante eventos específicos. Algoritmos genéticos, no fueron considerados en el estudio.

Se utilizó el Software Estadístico R, por su amplio contenido de librerías útiles para el proyecto y además es libre. Finalmente se procedió al acoplamiento de las diferentes herramientas.

RESULTADOS

Se obtuvo una aplicación informática amigable al usuario y que da respuestas con el debido sustento estadístico, realizado con los comandos de R. La misma ofrece, gráficos explicativos, editables y de fácil entendimiento con respecto a las variables seleccionadas, obtenidos inmediatamente ejecutada la consulta en el cubo de datos.



El análisis exploratorio de minería de datos, muestra tendencias en el tiempo y las localidades de las variedades estudiadas, clarificando la respuesta del genotipo en cada ambiente en el que se desarrolla.

Observadas las tendencias y haciendo uso de la Aplicación, se realizaron pruebas que validaran estadísticamente esos resultados; con la ayuda de las librerías del R se comprobó de manera numérica las diferencias apreciadas con su respectiva significancia estadística.

CONCLUSIONES

- Se detectaron patrones de respuesta agronómica y fitosanitarias de las variedades estudiadas, mostrando que las técnicas de minería de datos son aplicables en este tipo de estudios, siempre que se cuente con la data necesaria.

- El tiempo empleado en realizar la preparación de los datos, fue aproximadamente el 70% del tiempo del proyecto.

- Se aplicaron las tres técnicas de minería de datos más utilizadas, Árboles de decisión (56.6%), Agrupamiento (43.9%) y Estadística clásica (43.2%).

- La Aplicación desarrollada, describe cómo la variedad **de prueba** estudiada por el CIBE, evoluciona de manera muy favorable, en comparación con las demás. Además hubo discriminación entre las variedades, manteniéndose una estabilidad relativa en la variedad **de prueba**. La ubicación geográfica no afectó la estabilidad de la evolución en los parámetros agronómicos o fitosanitarios.

- Por la falta de datos de clima, quedaron sin respuesta preguntas muy importantes: ¿existe un patrón de crecimiento de algún parámetro relacionado directa o indirectamente con temperatura y/u otro parámetro climático?; ¿existe un patrón biológico entre el desarrollo de la Sigatoka versus mayores temperaturas y/u otro parámetro climático?