Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 50- 63

Tipo de artículo: Artículo original
Temática: Bioinformática
Recibido: 20/05/2017 | Aceptado: 25/06/2017

# Determinación heurística de las relaciones entre las estructuras y la producción de metabolitos secundarios

# *Heuristic determination of relations between the structures and the production of secondary metabolites*

**Jorge García Brizuela [1*], Cosme E. Santiesteban Toca [2], Yanelis Capdesuñer Ruíz [3]**

[1] Centro de Bioplantas, "Universidad Máximo Gomes Báez", Ciego de Ávila, Cuba, Carretera a Morón, Km 9½. CP: 69150. {jorgito,cosme}@bioplantas.cu
[2] Centro de Bioplantas, "Universidad Máximo Gómez Báez", Ciego de Ávila, Cuba, Carretera a Morón, Km 9½. CP. 69150. ycapdesuñer@bioplantas.cu
[*] Autor para correspondencia: jorgito@bioplantas.cu

---

**Resumen**

El cultivo y producción del tabaco (nicotiana tabacum), es altamente dependiente de los plaguicidas. La aplicación de plaguicidas a menudo no es eficaz y es peligrosa para los seres humanos y el medio ambiente. Sin embargo, es bien conocido que algunos metabolitos secundarios juegan un papel esencial en la protección de plantas contra patógenos. Sin embargo, establecer la relación entre la producción de metabolitos secundarios y los rasgos fenotípicos requiere una amplia experimentación y el seguimiento de la información manual. Además, desde el perfil fenotípico de una planta determinada podemos derivar la producción de determinados productos naturales, pero es imposible deducir el perfil fenotípico de las plantas en función de sus productos naturales. Por esta razón, el objetivo de esta investigación es diseñar un método basado en técnicas de aprendizaje automático, entrenado con los rasgos fenotípicos (morfología del tricomas) de las plantas, que sea capaz de aprender la correlación existente entre los rasgos fenotípicos y los metabolitos secundarios de las plantas de nicotiana tabacum. La búsqueda de la relación de los metabolitos con el nivel de expresión de un rasgo dado es un problema de regresión. Por lo que fueron empleados un grupo de técnicas de regresión basadas en estadígrafos tradicionales y técnicas de aprendizaje automático. Como resultado de la experimentación, se determinó que el empleo de un árbol de regresión REPTree permite determinar los rasgos que mejor correlacionan con el metabolito estudiado. Además, como valor agregado, es capaz de devolver un conjunto de reglas simples que describen este proceso.

**Palabras claves:** regresión, metabolitos, correlación, fenotipos.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 50- 63

**Abstract**

*The tobacco culture and production, especially in tropical countries, is highly dependent of pesticides. But, the application of pesticides is often ineffective and dangerous to the humans and the environment. Moreover, it is well known that some secondary metabolites play an essential role in the protection of plants against pathogens. However, establishing the relationship between the production of secondary metabolites and the phonotypical traits requires an extensive experimentation and manual tracking information. Moreover, from the metabolite profile of a particular plant we can derive its biological activity, but it is impossible to deduce the exact metabolic profile of plants according to their biological activity. Therefore, the objective of this research is to design a method based on machine learning techniques, trained on the plants phenotypic profile, which is able to learn the correlation between the secondary metabolites and phenotypic traits of tobacco plants. Finding the relationship between phenotypes and the expression level of a given metabolite, is a regression problem. For this reason, some traditional statistics techniques and machine learning techniques were employed. As a result of experimentation, it was determined that the use of a REPTree regression tree, determines the characteristics that best correlate with the studied metabolite. Also, as an added value, it is able to return a set of simple rules that describe this process.*

***Key Words:*** *correlation, metabolites, phenotypes, regression.*

---

## Introducción

En su hábitat natural las plantas deben enfrentar y protegerse de múltiples enemigos reales y potenciales, tales como plantas que crecen en su entorno, microorganismos que provocan enfermedades, insectos, entre otros. Además, deben competir con el resto de las plantas en su medio por agua, luz y nutrientes del suelo, entre otros tipos de estrés bióticos y abióticos (Bartwal et al. 2013).

En los organismos, el conjunto de reacciones químicas que tienen lugar constituyen el metabolismo. Donde, los metabolitos primarios están presentes en todas las plantas y desempeñan funciones similares. Mientras que los metabolitos secundarios son aquellos que no parecen tener una función directa en procesos fotosintéticos, respiratorios, entre otros, a los que también se denominan productos secundarios o productos naturales (Maag et al. 2015).

Estos metabolitos secundarios, además de no presentar una función definida en los procesos mencionados, difieren también de los metabolitos primarios ya que no todos se encuentran en todos los grupos de plantas y se sintetizan en pequeñas cantidades y no de forma generalizada, restringiéndose a un determinado género de plantas, familia, o especies.

Algunos productos del metabolismo secundario tienen funciones ecológicas específicas como atrayentes o repelentes de animales. También intervienen en los mecanismos de defensa de las plantas frente a diferentes patógenos, actuando como pesticidas naturales (Moore et al. 2014).

Sin embargo, establecer la relación que existe entre la producción de productos naturales, las estructuras morfológicas implicadas en el proceso y la actividad biológica, es una tarea que exige una extensiva experimentación y procesamiento de información, que hasta la actualidad se realiza de forma manual.

Esto se debe, en lo fundamental, a que de una estructura determinada se puede llegar a un metabolito y/o actividad biológica, sin embargo, a esa misma actividad biológica se puede llegar desde diferentes estructuras. La no existencia de una correspondencia directa y bidireccional implica que no se puede establecer con facilidad la relación que existe entre estructuras, metabolitos secundarios y actividad biológica (Maag et al. 2015).

Como estructuras se refiere a los pelos llamados tricomas por los que están cubiertas la mayoría de las plantas. Los tricomas son reconocidos dentro de las características anatómicas de las plantas o fenotipos (Kortbeek et al. 2016) y juegan un papel clave en la taxonomía de las plantas, se han definido como protuberancias epidérmicas que se distinguen por su tamaño y forma para su clasificación en diferentes tipos. Los tricomas se clasifican en dos grandes tipos: Tricomas no Glandulares o simples y tricomas Glandulares.

En la actualidad existen deficiencias en las técnicas para predecir la relación existente entre los rasgos fenotípicos y los metabolitos secundarios. El objetivo de la investigación es diseñar un método, basado en técnicas de aprendizaje automático y entrenado en el perfil fenotípico de las plantas, que sea capaz de aprender la correlación existente entre las estructuras y los productos naturales de las plantas de nicotiana tabacum.

## Materiales y métodos

### Métodos de selección de atributos

El proceso de selección de rasgos consta de dos componentes principales: una función de evaluación y un método de búsqueda (Figura 1). La función de evaluación permite calcular la calidad de un subconjunto de rasgos; mientras que el método de búsqueda, por lo general heurística, es el encargado de generar los subconjuntos de rasgos. Seleccionar los rasgos relevantes de un conjunto de datos es una tarea necesaria en el aprendizaje automatizado (*machine learning*), dada su importancia en el descubrimiento de reglas y relaciones en grandes volúmenes de datos entre otras aplicaciones (Wang et al. 2014).

El algoritmo "***CfsSubsetEval***" (CSE), calcula la correlación de la clase con cada atributo, y eliminan atributos que tienen una correlación muy alta como atributos redundantes. Los subconjuntos de características que están altamente correlacionadas con la clase al tener inter-correlación baja son preferidos (Park and Hong 2014).

En cuanto al método de búsqueda, vamos a mencionar el "***BestFirst***" (BF), que recorre el espacio de subconjuntos de atributos por medio del método escalador de montaña ávido, engrandecido con una facilidad que vuelve hacia atrás (backtracking). El método "***GeneticSearch***" (GS) realiza una búsqueda usando el simple algoritmo genético descrito en (De'ath 2007).

## Técnicas de predicción

Los algoritmos **supervisados o predictivos** predicen el valor de un atributo (clase) de un conjunto de datos, conocidos otros atributos (rasgos). A partir de datos cuya clase se conoce, por tanto, se induce una relación entre dicha clase y los rasgos, el cual se desarrolla en dos fases:

Entrenamiento (construcción de un modelo usando un subconjunto de datos con clase conocida) y prueba (prueba del modelo sobre el resto de los datos). Existen diferentes tipos de algoritmos de aprendizaje automático. El algoritmo **LinearRegression** (LR) **e**s el más sencillo para hacer predicción realizando regresión lineal. Usa como criterio de selección del modelo Akaike(Akaike 1974), y puede manejar datos incompletos. El **DecisionTable (**DT) es un algoritmo que más que un árbol, la tabla de decisión es una matriz de filas y columnas que indican condiciones y acciones. Las reglas de decisión, incluidas en una tabla de decisión, establecen el procedimiento a seguir cuando existen ciertas condiciones.

El algoritmo **M5P** se basa en árboles de decisión, sin embargo, en lugar de tener valores en los nodos de árbol, contiene un modelo lineal multi-variable de regresión en cada nodo. El espacio de entrada está dividido en celdas de entrenamiento de datos y sus salidas, luego se crea un modelo de regresión en cada celda como una hoja del árbol(Quinlan 1992). El **M5R** genera una lista de decisiones para los problemas de regresión usando el paradigma divide y vencerás. En cada iteración construye un árbol modelo utilizando el algoritmo M5 y convierte la mejor hoja en una regla (Azofra et al. 2015). Este algoritmo tiene el mismo principio de funcionamiento que el M5P. Lo que, en vez de construir un modelo en forma de árbol, describe reglas. Pese a su funcionamiento muy parecido, devuelve resultados muy semejantes, pero nunca iguales.

El **REPTree** (RT) es un método de aprendizaje rápido mediante árboles de decisión. Construye un árbol de decisión usando la información de varianza y lo poda usando como criterio la reducción del error. Solamente clasifica valores

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 50- 63

para atributos numéricos una vez. Los valores ausentes se manejan dividiendo las instancias correspondientes en segmentos. Es un árbol de clasificación con modelo comprensible (reglas *if then else*) (Jung, Kang, and Choi 2016).

## Herramienta utilizada

La herramienta utilizada en la investigación es la plataforma experimental Weka (*Waikato Environment for Knowledge Analysis*, Entorno para Análisis del Conocimiento de la Universidad de Waikato). Es una plataforma de software para resolver problemas de aprendizaje automático y minería de datos. Es escrito en Java y desarrollado en la Universidad de Waikato. Weka es un software libre distribuido bajo licencia GNU-GPL

## Preparación de la muestra y procedimiento de cuantificación:

Material vegetal: Para el estudio de la morfología de los tricomas se tomaron como muestra 2 hojas (Hoja joven completamente desplegada de aproximadamente 15 a 18 cm de largo en la segunda posición en el tallo después de las hojas pequeñas en formación, Figura ) de cada planta por línea de cultivo, de las 8 líneas de cultivo de *nicotiana tabacum* empleadas para la extracción del material vegetal. Las plantas utilizadas para la investigación en el momento de extracción de las muestras tenían 3 meses de edad crecidas en condiciones de campo. De ellas se realizaron preparaciones fijas para observar y fotografiar en el microscopio electrónico de forma tal que se obtuvieron un total de 9 imágenes de diferentes zonas del haz y 9 imágenes de diferentes zonas del envés por línea.

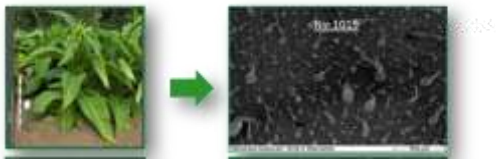

Figura 1. Extracción del material vegetal.

A la izquierda planta de 3 meses de edad en condiciones de campo y a la derecha imágenes tomadas por el microscopio electrónico de barrido, tanto del haz como del envés de la hoja.

## Clasificación de los tipos encontrados:

Se clasificaron los tricomas glandulares según la clasificación establecida en *nicotiana tabacum* de dos tipos (Tipo I y Tipo II) pero además se hicieron otras clasificaciones según la morfología observada. La clasificación se realizó manual para las 18 imágenes con colores definidos en Photoshop para cada tipo de tricoma (Figura ).

Los Tipo I G marcados en rojo son tallos largos pluricelulares y glándula con un tamaño ≥300 µm, Tipo I R en Verde son tallos largos pluricelulares ramificados, Tipo I P en Azul son tallos largos de pluricelulares generalmente pequeños

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 50- 63

con un tamaño < 300 µm y Tipo II en Amarillo identificados como tallos cortos de glándulas esféricas uni o pluricelulares.

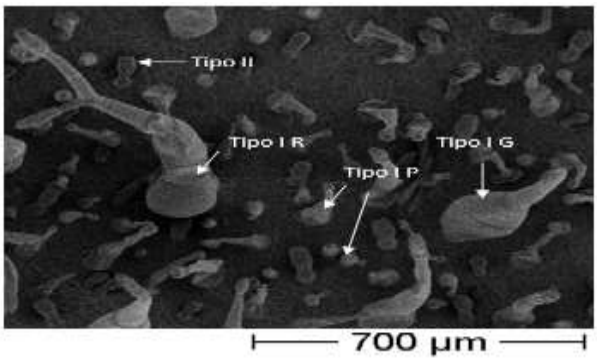


Figura 2. Clasificación de los tricomas.

Realizada para el conteo y análisis de la morfología, y distribución de los tricomas en el haz y el envés de hojas de 8 líneas de nicotiana tabacum con perfiles metabólicos contrastantes de sus exudados foliares. TipoI G: Grandes, TipoI R: Ramificado, Tipo I P: Pequeño y Tipo II.

El conteo de tricomas se realizó con el empleo de un software desarrollado por el Doctor Milton García del departamento de Bioinformática del Centro de Bioplantas para el conteo por colores de tricomas, luego de su identificación y clasificación previa con el empleo de Photoshop. Se evalúan las siguientes variables el total de tricomas por línea, el total de tricomas en el haz, el total de tricomas en el envés, y el total de tricomas por tipo tanto en el haz como el envés de la hoja (Figura ).

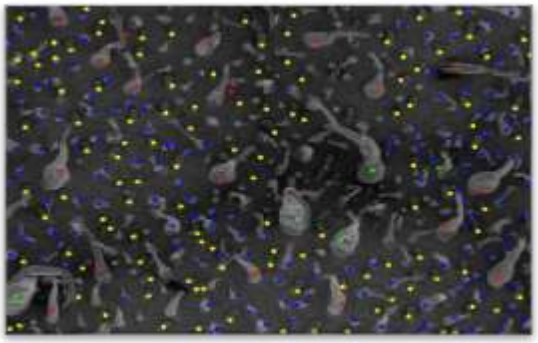

Figura 3. Conteo de tricomas glandulares por colores según sus tipos y morfologías observadas.

La imagen muestra los tipos de tricomas clasificados por colores, para después ser contabilizados por la herramienta pertinente. Este proceso es realizado por un especialista en la materia.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 50- 63

**Bases de datos**

Para la experimentación en el método a proponer, fueron empleados seis conjuntos de datos, uno por cada producto natural:

**1. Ds-ACd**: Producto Natural Alpha_CBT_diol
**2. Ds-ACo**: Producto Natural Alpha_CBT_ol
**3. Ds-BCd**: Producto Natural Beta_CBT_diol
**4. Ds-BCo**: Producto Natural Beta_CBT_ol
**5. Ds-Ca**: Producto Natural Cis_abienol
**6. Ds-Ld**: Producto Natural Labdenediol

Cada conjunto contiene 72 objetos y estos tienen quince diferentes mediciones de rasgos fenotípicos y una clase. Las mediciones y la clase presentas datos numéricos y no presentan datos incompletos. Las base de datos fueron codificadas en formato *.arff por su fácil manejo y ser el tipo de fichero de entrada por defecto de la plataforma de experimentación Weka.

**Codificación del vector de entrada**

El vector de entrada es de longitud fija, incluye la información de los rasgos fenotípicos (atributos) y la medición de un compuesto (clase) de una planta. El vector de entrada contiene 15 rasgos y una clase. Los rasgos se dividen en tres grupos de cinco. El grupo uno y dos corresponde al haz y el envés de la hoja respectivamente. En estos dos grupos cuatro de los rasgos corresponde al conteo de un tipo de tricoma específico, que es común en el haz y el envés, y el quinto rasgo es la suma de los cuatro anteriores.

En el tercer grupo los cuatro primeros atributos son el total de tricomas por tipos y el quinto atributo es la suma de los totales de los dos grupos anteriores (Figura ). La clase hace referencia a la medición de uno de los metabolitos secundario estudiados.

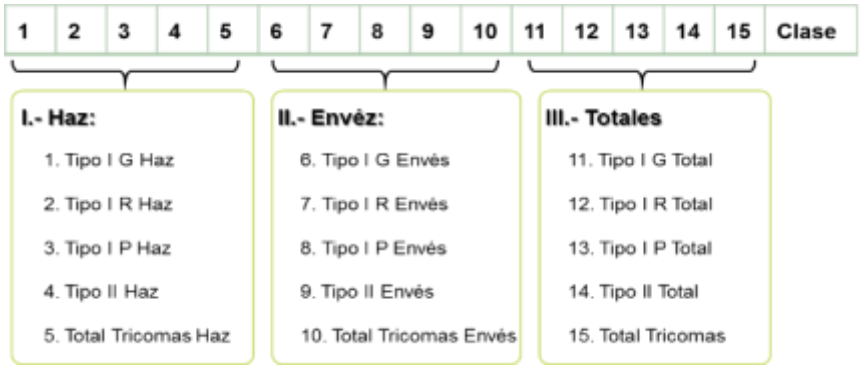


Figura 4. Codificación del vector de entrada.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 50- 63

La figura muestra el vector de entrada dividido en tres, más una clase de naturaleza numérica. La primera y segunda parte se hacen corresponder a los tricomas por tipo más el total del haz y el envés de la hoja, respectivamente. En la tercera parte aparecen los totales por tipos tanto del haz como del envés y el total general de tricomas.

# Resultados y discusión

## Selección de los fenotipos empleando aprendizaje automático

La selección de los rasgos fenotípicos se realizó empleando los métodos de aprendizaje automático ***"CfsSubsetEval + BestFirst"*** (CSE-Bf) *y* ***"CfsSubsetEval + GeneticSearch"*** (CSE-Gs). Tanto CSE-Bf como CSE-Gs arrojaron la misma selección de rasgos para los seis conjuntos de datos empleados (Tabla ).

Tabla 1. Selección de rasgos fenotípicos aplicando aprendizaje automático.

| C. de Datos | Rasgos |
|---|---|
| **Ds-ACd** | Tipo I G Haz, Tipo I R Envés, Tipo II Envés, **Total Tricomas Envés** |
| **Ds-ACo** | Tipo I R Envés, **Total Tricomas Envés**, Tipo I R Total |
| **Ds-BCd** | Tipo I G Haz, Tipo II Envés, **Total Tricomas Envés**, Tipo I G Total, Tipo II Total |
| **Ds-BCo** | Tipo I R Haz, Tipo I G Envés, **Total Tricomas Envés**, Tipo I R Total |
| **Ds-Ca** | Tipo II Haz, Tipo I P Envés |
| **Ds-Ld** | Tipo_I_P_Envés |

En la primera columna aparecen los conjuntos de datos a los que le fue aplicado los métodos de aprendizaje automático. En la segunda columna la selección de rasgos hecha por los algoritmos CSE-Bf y CSE-Gs.

Como se pudo observar en la tabla los rasgos seleccionados en su gran mayoría pertenecen al envés de la hoja de *nicotiana tabacum*. También se aprecia que el rasgo **"Total Tricomas Envés"** fue elegido en 4 de los 6 conjuntos de datos. Estos resultados difieren por completo de los obtenidos por las técnicas estadísticas. Sin embargo, el alto nivel de correlación y la inclusión del nivel de expresión de los metabolitos estudiados en el modelo, dan crédito a estos resultados.

## Determinación de la relación entre los rasgos y la producción de los metabolitos estudiados

Para hacer el análisis de la relación rasgos-producción se emplearon los siguientes algoritmos de regresión:

- **LR:** este algoritmo fue configurado con un algoritmo goloso como método de selección de atributos (*greedy*) y valor de $\mathbf{1 \times 10^{-10}}$ en la cordillera de selección ("-S 2 -R 1.0E-8").

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 50- 63

- **DT:** la configuración para este algoritmo se hizo aplicándole un valor de 1 en la validación cruzada. Como método de selección de atributos BF con sus valores por defecto ("-X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5").
- **M5R y M5P:** estos algoritmos fueron configurados para aceptar un mínimo de cuatro objetos por hojas ("-M 4.0")
- **RT:** este último algoritmo fue configurado con un mínimo de 2 objetos por hoja, una varianza promedio mínima de 0.001. Para la poda se especificó un valor 5 particiones, para dar aleatoriedad agrego un 4 como valor en la semilla y -1 como representación de la no restricción en la profundidad del árbol construido ("-M 2 -V 0.001 -N 5 -S 4 -L -1").

Donde, cada una de las configuraciones aplicadas a estos fueron obtenidas a partir de un proceso de análisis de la bondad de ajuste. Posteriormente, estos algoritmos fueron evaluados empleando un procedimiento de validación cruzada con 10 particiones. Los resultados experimentales se muestran en la Tabla , donde, cada algoritmo fue analizado para cada conjunto de datos por separado. Tomando en cuenta como medida el CC para predecir la relación entre los rasgos fenotípicos y la producción de metabolitos secundarios.

Tabla 2. Comportamiento de los algoritmos predictivos en base al CC.

| | LR | DT | RT | M5P | M5R |
|---|---|---|---|---|---|
| **Ds-ACd** | 0,7232 | 0,6861 | 0,8612 | **0,8663** | 0,8486 |
| **Ds-ACo** | 0,3865 | 0,7647 | **0,9151** | 0,8415 | 0,8959 |
| **Ds-BCd** | 0,7877 | 0,6275 | **0,8695** | 0,7593 | 0,805 |
| **Ds-BCo** | 0,3893 | 0,7727 | **0,8825** | 0,8712 | 0,8463 |
| **Ds-Ca** | 0,53 | 0,5594 | **0,9063** | 0,8389 | 0,8279 |
| **Ds-Ld** | 0,4237 | 0,6602 | **0,8908** | 0,768 | 0,8376 |

Los algoritmos fueron evaluados empleando un procedimiento de validación cruzada con 10 particiones. Cada algoritmo fue analizado para cada conjunto de datos por separado. Tomando en cuenta como medida el CC para predecir la relación de los rasgos fenotípicos con la producción de metabolitos secundarios.

El análisis de esta tabla muestra el comportamiento de los 5 algoritmos en los 6 conjuntos de datos. En general los algoritmos RT, M5P y M5R muestran un comportamiento estable en los diferentes conjuntos, superior a 0,75. El algoritmo RT fue el de mejor desempeño de todos con valores entre 0,86 y 0,92. Los algoritmos LR y DT son los de menor desempeño y mayor variación en el CC, con valores que van desde 0,38 a 0,79 (Figura ).

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 50- 63

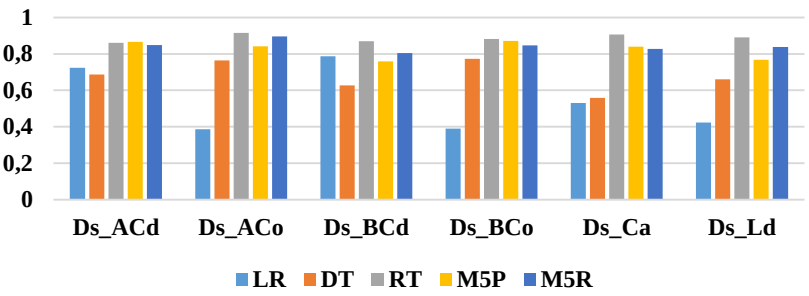


Figura 4. Comportamiento de los algoritmos en base al CC.

En el eje de abscisas se encuentran los algoritmos predictivos aplicados agrupados por conjunto de datos. En el eje de ordenadas, la medición del CC.

La figura afirma de forma gráfica la paridad entre RT, M5P y M5R. Evidenciando una ligera ventaja de RT sobre M5P y M5R, respectivamente. Tras el análisis se aprecia que este algoritmo presenta el mejor desempeño en 5 de las 6 bases de datos. Sin embargo, estas valoraciones son muy dependientes de la escala, por lo que será necesario aplicar y analizar los test significación estadística en los resultados.

## Aplicación de test estadísticos

Sin embargo, el análisis de una gráfica no resulta suficiente para poder determinar si existen realmente diferencias en el comportamiento entre los algoritmos analizados. Con el objetivo de poder refutar esta hipótesis, resulta aconsejable la consideración de algún test estadístico.

Con el objetivo de determinar si existen diferencias estadísticamente significativas en los resultados de los algoritmos para los diferentes conjuntos de datos, fueron empleados múltiples pruebas estadísticas no paramétricas. Mediante la prueba de Friedman se comprueba la hipótesis nula de que los algoritmos obtienen resultados similares [15], como promedio, para todos los conjuntos. Al ser rechazada la hipótesis nula, entonces fueron aplicadas las pruebas de *post-hoc*: Bonferroni-Dunn, Holm, Hochberg y Hommel. Para los niveles de significación de 0.05 y 0,1 (Tabla ).

Tabla 3. Resultados del test de Friedman.

| ***Algoritmo*** | ***Ranking*** |
|---|---|
| *REPTree* | 1,2 |
| *M5P* | 2,5 |
| *M5R* | 2,5 |
| *DecisionTable* | 4,3 |
| *LinearRegression* | 4,5 |

Considerando una reducción del desempeño distribuida (acorde a un shi cuadrado con 4 grados de libertad): 18,9333 y un p-value = 0,0008.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 50- 63

El resultado de la aplicación del test de Friedman rechaza la hipótesis nula, arrojando que si existen diferencias estadísticamente significativas. Primero el algoritmo RT teniendo un comportamiento superior, seguido de M5R y M5P donde estos dos no presentan diferencias significativas entre ellos. Por último, aparecen DT y LR por ese orden respectivamente. En el **¡Error! No se encuentra el origen de la referencia.** aparecen los resultados de los test estadísticos en su totalidad para valores de significación de 0,05 y 0,10.

## Mecanismo de interpretación

El algoritmo RT, al estar basado en árboles de regresión, tiene la habilidad de poseer un modelo de conocimiento interpretable. Lo cual le proporciona una apreciable ventaja sobre el resto de los algoritmos basados en redes neuronales, modelos probabilísticos, máquinas de vectores soporte, entre otros. Debido a que las bases de conocimientos de las redes neuronales son matrices de pesos o en las máquinas de soporte de vectores son matrices de vectores, las cuales no pueden ser explicadas con facilidad por el biólogo (usuario final).

Como resultado final se propone un árbol de regresión (RT) por cada uno de los 6 conjuntos de datos. Donde, cada hoja muestra la cantidad del metabolito a producir con un nivel de confianza que está en dependencia de la cantidad de veces que se observa el patrón. Cada árbol se construye a partir de diferentes conjuntos de datos, lo cual condiciona la diversidad entre ellos. Esto se puede apreciar en la Figura , la cual muestra el árbol de regresión construido para el conjunto de datos Ds-ACd, pero solo con los rasgos fenotípicos seleccionados por los algoritmos de selección de atributos CSE-Gs y CSE-Bf (Tabla ).

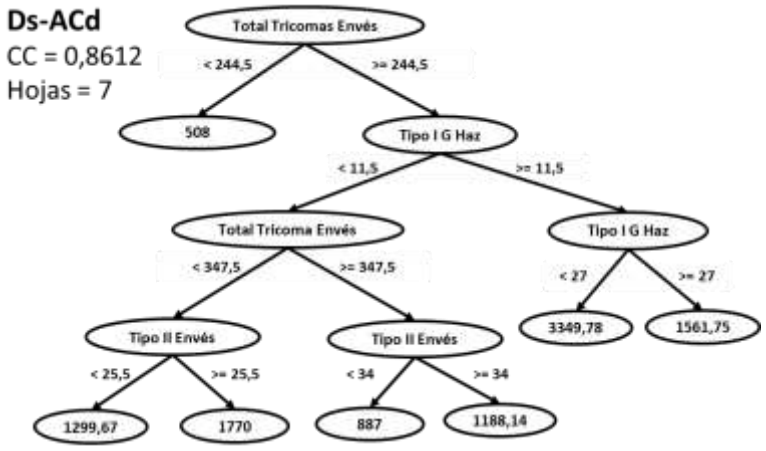


Figura 5. Árbol de predicción generado por RT a partir del conjunto de datos Ds-ACd.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 50- 63

Como se observa en la figura el árbol modelo describe la forma en que se comportará la producción del metabolito Ds-ACd a partir de los rasgos que mejor lo predicen, con una alta confianza del 0,8612. El camino desde la raíz a cada una de las 7 hojas serían las reglas que describan dicha producción, como se muestra:

- ***Si*** Total Tricomas Envés < 244,5 → Ds-ACd = 508
- ***Si*** Total Tricomas Envés >= 244,5 ***y*** Tipo I G Haz < 11,5 y Total Tricomas Envés < 347,5 ***y*** Tipo II Envés < 25,5 → Ds-ACd = 1299,67
- ***Si*** Total Tricomas Envés >= 244,5 ***y*** Tipo I G Haz < 11,5 y Total Tricomas Envés < 347,5 ***y*** Tipo II Envés >= 25,5 → Ds-ACd = 1770
- ***Si*** Total Tricomas Envés >= 244,5 ***y*** Tipo I G Haz < 11,5 y Total Tricomas Envés < 347,5 ***y*** Tipo II Envés < 34 → Ds-ACd = 887
- ***Si*** Total Tricomas Envés >= 244,5 ***y*** Tipo I G Haz < 11,5 y Total Tricomas Envés < 347,5 ***y*** Tipo II Envés >= 34 → Ds-ACd = 1188,14
- ***Si*** Total Tricomas Envés >= 244,5 ***y*** Tipo I G Haz >= 11,5 y Tipo I G Haz < 27 → Ds-ACd = 3349,78
- ***Si*** Total Tricomas Envés >= 244,5 ***y*** Tipo I G Haz >= 11,5 y Tipo I G Haz >= 27 → Ds-ACd = 1561,75

La capacidad de RT de generar una pequeña colección de reglas con una alta confianza representa una gran ventaja para su mecanismo de interpretación. RT brinda el conjunto de reglas ordenadas en cuanto al nivel de confianza y cobertura. Esto facilita el trabajo de los investigadores, responsables de "descubrir" indicios de cómo se realiza la producción de los metabolitos, dando explicación biológica de alguno de dichos supuestos (reglas).

## Conclusiones

A partir del empleo de un conjunto de técnicas heurísticas, fueron determinados los rasgos fenotípicos que guardan relación directa con la producción de cada metabolito estudiado. Logrando un alto nivel de correlación y precisión, en comparación con los métodos estadísticos tradicionales. La extracción de estos rasgos facilita la caracterización de las plantas Nicotiana tabacum, en cuanto a la capacidad de producir estos compuestos naturales.

Se implementó un método, basado en los árboles de decisión REPTree, el cual es capaz de aprender la correlación entre rasgos fenotípicos y metabolitos secundarios de Nicotiana tabacum con un coeficiente de correlación desde el 86% para Alpha_CBT_diol hasta el 91% para Alpha_CBT_ol. El cual tiene, como valor agregado, la capacidad de generar modelos entendibles para los investigadores.

## Referencias

AKAIKE, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, pp. 716-723.

AZOFRA, D., MARTÍNEZ, E., JIMÉNEZ, E., BLANCO, J., AZOFRA, F. y SAENZ-DÍEZ, J.C., 2015. Comparison of the influence of photovoltaic and wind power on the Spanish electricity prices by means of artificial intelligence techinques. *Renewable and Sustainable Energy Reviews* [en línea], vol. 42, pp. 532-542. ISSN 13640321. DOI 10.1016/j.rser.2014.10.048. Disponible en: http://dx.doi.org/10.1016/j.rser.2014.10.048.

BARTWAL, A., MALL, R., LOHANI, P., GURU, S.K. y ARORA, S., 2013. Role of Secondary Metabolites and Brassinosteroids in Plant Defense Against Environmental Stresses. *Journal of Plant Growth Regulation*, vol. 32, no. 1, pp. 216-232. ISSN 07217595. DOI 10.1007/s00344-012-9272-x.

DE'ATH, G., 2007. Boosted regression trees for ecological modeling and prediction. *Ecology* [en línea], vol. 88, no. 1, pp. 243-251. ISSN 00129658. DOI 10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2. Disponible en: http://www.ncbi.nlm.nih.gov/pubmed/17489472.

JUNG, Y.G., KANG, M.S. y CHOI, Y.J., 2016. *Using J48 and REPTree to predict risk factors in medicine*. 2016. S.l.: s.n.

KORTBEEK, R.W.J., XU, J., RAMIREZ, A., SPYROPOULOU, E., DIERGAARDE, P., OTTEN-BRUGGEMAN, I., DE BOTH, M., NAGEL, R., SCHMIDT, A., SCHUURINK, R.C. y BLEEKER, P.M., 2016. Engineering of Tomato Glandular Trichomes for the Production of Specialized Metabolites. *Methods in Enzymology*, ISSN 15577988. DOI 10.1016/bs.mie.2016.02.014.

MAAG, D., ERB, M., KÖLLNER, T.G. y GERSHENZON, J., 2015. Defensive weapons and defense signals in plants: Some metabolites serve both roles. *BioEssays*, vol. 37, no. 2, pp. 167-174. ISSN 15211878. DOI 10.1002/bies.201400124.

MOORE, B., ANDREW, R., KÜLHEIM, C. y FOLEY, W., 2014. Explaining intraspecific diversity in plant secondary metabolites in an ecological context. *New Phytologist* [en línea], vol. 201, pp. 733-750. Disponible en: http://onlinelibrary.wiley.com/doi/10.1111/nph.12526/full.

PARK, M. y HONG, E., 2014. Software Fault Prediction Model using Clustering Algorithms Determining the Number of Clusters Automatically. *International Journal of Software Engineering & Its …* [en línea], vol. 8, no. 7, pp. 199-204.

Disponible en: http://www.sersc.org/journals/IJSEIA/vol8_no7_2014/16.pdf.

QUINLAN, R.J., 1992. Learning with Continuous Classes. *5th Australian Joint Conference on Artificial Intelligence*. Singapore: s.n., pp. 343-348.

WANG, C., HE, Q., CHEN, D. y HU, Q., 2014. A novel method for attribute reduction of covering decision systems. *Information Sciences* [en línea], vol. 254, pp. 181-196. ISSN 00200255. DOI 10.1016/j.ins.2013.08.057. Disponible en: http://dx.doi.org/10.1016/j.ins.2013.08.057.