Original

Aplicación de la minería de datos en la estimación de componentes fotoquímicos

Application of data mining in the estimation of photochemical components

Ing. Pedro Manuel Estrada Jiménez, Facultad de Educación Media, Universidad de Granma, Cuba, pestradaj@udg.co.cu

Dr. C. Jorge Luis Ramírez de la Ribera, Centro de Estudio de Producción Animal, Universidad de Granma, Cuba, jramirezrivera@udg.co.cu

Dr. C. Danis Manuel Verdecia Acosta. Centro de Estudio de Producción Animal, Universidad de Granma, Cuba, dverdeciaa@udg.co.cu

Dr. C. Yolanda Soler Pellicer, CIGET Granma, Cuba,

ysolerp@ciget.granma.inf.cu

Recibido: 23/11/2018 Aceptado: 12/04/2019

Resumen

Con el objetivo de simular el comportamiento de las variedades *Leucaena leucocephala* y *Tithonia diversifolia* se emplearon técnicas de minería de datos para la determinación de un modelo de predicciones. Para ello se usó una base de datos proporcionada por el personal científico perteneciente al Centro de Estudios de Producción Animal (CEPA) de la Universidad de Granma. Estos valores se obtuvieron mediante un análisis a las muestras fuera del país, donde se estudió la cuantificación de los componentes fitoquímicos. Las variedades fueron cultivadas en el Valle del Cauto de la provincia de Granma. Para cada variedad se registraron los metabolitos primarios, variables de clima y los metabolitos secundarios para cada uno de esos valores. Se probaron los clasificadores de weka en busca del clasificador que reportara menor error cuadrático medio para conformar con este resultado un modelo de aprendizaje multietiquetas. Los experimentos a los modelos determinados estuvieron compuestos por valores con los que el sistema no fue entrenado para determinar el nivel de certeza de las predicciones.

Palabras clave: metabolitos secundarios; composición fitoquímica; clasificación; multietiqueta.

Abstract

In order to simulate the behavior of the varieties Leucaena Leucocephala and Tithonia diversifolia, data mining techniques were used to determine a prediction model. This was used a database provided by the scientific staff belonging to the Center for Animal Production Studies

(CEPA) of the University of Granma. These values were obtained by means of an analysis to the samples outside the country, where the quantification of the phytochemicals components was studied. The varieties were cultivated in the cautious valley of Granma province. For each variety, primary metabolites, climate variables and secondary metabolites were recorded for each of these values. We tested the classifiers of Weka in search of the classifier to report lower quadratic error mean to form with this result a Multilabel learning model. The experiments to the specific models were composed of values with which the system was not trained to determine the level of certainty of the predictions.

Keywords: Secondary metabolites, phytochemical composition, classification, multi label.

Introducción

La productividad de las plantas está determinada por un grupo de factores inherentes al vegetal y externos. En el primero se encuentran sus características biológicas y el segundo, el suelo, clima y manejo (Herrera *et al*. 2017).Estos aspectos cobran en la actualidad particular importancia, debido al elevado precio de las materias primas para la producción de alimentos concentrados y de los fertilizantes (Friedrich 2014).

De acuerdo con (Domínguez Gómez et al., 2012) los árboles y arbustos contribuyen a asegurar una dieta nutritiva para el ganado. Muchos árboles forrajeros tienen su hábitat en zonas áridas y semiáridas donde las condiciones del medio son difíciles para el cultivo de pastos introducidos; en estos casos, el pastoreo depende exclusivamente de dichos árboles y arbustos ya que, de otra forma, los animales no sobrevivirían. En numerosas regiones del mundo, la mayor parte de los ganaderos y pastores reconocen el valor de los árboles y han establecido medidas para su protección; pero en otras, estas medidas se las dejan a la naturaleza que actúa sobre la misma regeneración de las especies.

En el trópico la principal fuente de nutrientes y la más barata, para la alimentación del ganado vacuno la constituyen los pastos y forrajes, lo que se apoya en su economía y en la no competencia con las necesidades de alimentos para el consumo humano directo y de otros animales. Sin embargo, su crecimiento y productividad está influida por las condiciones climáticas existentes principalmente por la distribución anual de las lluvias, que unido a otros factores del medio ambiente y de manejo, repercuten en que estos no reflejen totalmente su potencialidad productiva y nutritiva (Vega Espinosa, Ramírez De la Ribera, Leonard Acosta, & Igarza, 2006)

Son varias las plantas forrajeras y arbustivas destinadas a la alimentación animal y entre ellas se encuentran *Leucaena leucocephala* y *Tithonia diversifolia*. De acuerdo con (Herrera, Verdecia, Ramírez, García, & Cruz, 2017) la *Tithonia diversifolia* es una planta de gran plasticidad ecológica con potencial forrajero para la alimentación y producción del ganado vacuno con aceptable calidad y alto valor proteico pero su composición química es variable.

Según (Galindo et al., 2014) en los últimos diez años ha cobrado gran importancia el uso de árboles y arbustos de leguminosas u otras especies, como suplemento para la dieta de los animales rumiantes. Estas plantas poseen características, como la presencia de metabolitos secundarios, que hacen que sean muy valoradas. Los metabolitos secundarios pueden modificar la velocidad de degradación y pasaje de los nutrientes a través del tracto gastrointestinal como resultado de un efecto directo en la ecología ruminal.

Teniendo en cuenta que los metabolitos secundarios, son un mecanismo de defensa de las plantas hay que estudiar los niveles de los mismos en las plantas que pueden ser usadas para alimento animal como plantea (Sepúlveda Jiménez, Porta Ducoing, & Rocha Sosa, 2003). (Carmona Agudelo, 2007) refiere que una gran cantidad de leguminosas arbóreas y arbustivas tropicales contienen factores o componentes anti nutricionales como los taninos, saponinas y otros que dificultan la digestibilidad en los animales, en especial los taninos. Estudios realizados a los metabolitos secundarios han aportado valores positivos en cuanto a su investigación como los expresados por (García, Ojeda, & Montejo, 2003) donde se realiza un estudio detallado de las concentraciones de algunos en distintas edades de cuatro variedades.

Al realizar el estudio para determinar la cuantificación de los metabolitos secundarios a una planta, primeramente se cosecha y al cabo de un tiempo (edad de rebote) realizarle un corte y ponerla en una estufa a aproximadamente 60 ^{o} por un período de 72 horas para obtener la materia seca. Esta materia es llevada a un laboratorio donde exista equipamiento y reactivos para determinar estos componentes y luego realizar los procedimientos pertinentes En la actualidad, estos estudios cuestan cada uno, alrededor de 60 €, por tanto por cada corte que se le realice a una planta hay que multiplicar el costo de aplicación de técnicas para determinar estos componentes.

El objetivo de este trabajo fue el entrenamiento de un clasificador para estimar los metabolitos secundarios de las variedades *Leucaena leucocephala* y *Tithonia diversifolia*.

Población y muestra

El proceso de determinación de metabolitos secundarios pasa en la actualidad por varias fases, la primera es la plantación de los arbustos, la cual se tuvo concebida en el valle del cauto, luego

se realizó un corte a una de las plantas para determinar (cualificar) los componentes fitoquímicos; para esto se empleó el tamizaje fitoquímico, realizado en la Universidad de Granma, estos componentes se muestran en la tabla 1. Una vez identificados los metabolitos secundarios presentes en las variedades se procedió al corte en diferentes edades de las plantas, se tomaron muestras en períodos de 60, 120 y 180 días en las etapas de lluvia y poca lluvia; seguidamente fueron depositadas las muestras en una estufa donde se mantuvieron a 60^{o} por un período de 72 horas con el objetivo de deshidratarlas y obtener la materia seca.

Tabla 1. Metabolitos secundarios de las variedades Leucaena leucocephala y Tithonia diversifolia.

Metabolitos secundarios
Taninos Totales
Taninos Condensados Totales
Taninos Condensados Ligados
Taninos Condensados Libres
Fenoles Totales
Verbascosa
Estaquiosa
Rafinosa
Flavonoides
Alcaloides
Saponinas
Triterpenos
Esteroides

A partir del paso anterior se realizaron los trámites convenientes para llevar estas plantas a un centro donde pudiera determinarse la cuantificación de estos metabolitos a las muestras tomadas; para este procedimiento se emplearon las técnicas tradicionales a partir de la aplicación de reactivos a las mismas. Una vez terminado este proceso de determinación de los componentes se construyó una base de datos con los valores de los metabolitos primarios y secundarios de las variedades *Leucaena leucocephala* y *Tithonia diversifolia* en un período de dos años, estos datos obtenidos en laboratorio todos están dados en el orden de los números reales, por lo tanto $x \epsilon \mathbb{R} / x > 0$.

Una vez terminado todo el proceso de obtención de los componentes fitoquímicos se determinaron las variables de entrada y variables de salida para la determinación de un modelo

computacional que haciendo uso de técnicas de inteligencia artificial pudiera predecir los valores de estos componentes; la identificación de dichas variables se muestra en la tabla 2 donde todas pertenecen al dominio de los números reales.

Tabla 2. Relación de variables creadas para cada uno de los parámetros a evaluar.

	Variable	Código	Dominio $\mathbf{x}\epsilon\mathbb{R}/\mathbf{x} > 0$
Variables de entrada	Variedad	V	$\mathbb{R}^+$
	Período	P	$\mathbb{R}^+$
	Edad	E	$\mathbb{R}^+$
	Nitrógeno	N	$\mathbb{R}^+$
	Glucosa	Gl	$\mathbb{R}^+$
	Fructosa	Fr	$\mathbb{R}^+$
	Sacarosa	Sc	$\mathbb{R}^+$
	Temperatura Máxima	TMax	$\mathbb{R}^+$
	Temperatura Mínima	TMin	$\mathbb{R}^+$
	Temperatura Media	TMed	$\mathbb{R}^+$
	Humedad Relativa Máxima	HRMax	$\mathbb{R}^+$
	Humedad Relativa Mínima	HRMin	$\mathbb{R}^+$
	Humedad Relativa Media	HRMed	$\mathbb{R}^+$
	Lluvia	Lll	$\mathbb{R}^+$
	Días con Lluvia	Dll	$\mathbb{R}^+$
Variables de salida	Taninos Totales	Tt	$\mathbb{R}^+$
	Taninos Condensados	Tct	$\mathbb{R}^+$
	Taninos Condensados	Tclt	$\mathbb{R}^+$
	Taninos Condensados	Tcl	$\mathbb{R}^+$
	Fenoles Totales	Ft	$\mathbb{R}^+$
	Verbascosa	Vr	$\mathbb{R}^+$
	Estaquiosa	Es	$\mathbb{R}^+$
	Rafinosa	Rf	$\mathbb{R}^+$
	Flavonoides	Fl	$\mathbb{R}^+$
	Alcaloides	Al	$\mathbb{R}^+$
	Saponinas	Sp	$\mathbb{R}^+$
	Triterpenos	Tr	$\mathbb{R}^+$
	Esteroides	Et	$\mathbb{R}^+$

Análisis de los resultados

A partir de conocer la cuantificación de los metabolitos secundarios se probaron clasificadores de la herramienta weka en busca del clasificador que reportara menor error cuadrático medio para la conformación de un sistema para predecir estos componentes partiendo del aprendizaje

de dicho sistema mediante la base de datos obtenida. Los resultados del aprendizaje de estos clasificadores se muestran en la tabla 3.

Tabla 3. Pruebas con clasificadores.

Clasificador	RMSE ***Leucaena***	RMSE ***Tithonia***
DecisionStump	2,8875±0,432	0,9383±0,068
MultilayerPercep	0,1505±0,032	0,0945±0,013
GaussianProces	1,6809±0,315	0,2693±0,045
SMOreg	1,5003±0,747	0,1481±0,075
M5P	0,7477±0,135	0,1789±0,026
REPTree	0,4527±0,534	0,1469±0,134
LinearRegressio	1,0690±0,207	0,1466±0,052
IBk	0,1176±0,026	0,0814±0,027
KStar	**0,0933±0,023**	**0,0783±0,022**
LWL	0,8819±0,128	0,3989±0,074

Como se observa el clasificador que resultó con mejor resultado fue el KStar. Según (Figueroa, Manríquez, Melesio, Mendoza, & Pérez, 2018) K * es un clasificador basado en instancias, que es la clase de una instancia de prueba se basa en la clase de esas instancias de capacitación similares a la misma, según lo determinado por una función de similitud. Se diferencia de otros estudios basados en instancia en que utiliza una función de la distancia basada en la entropía. Es un método en el que se aplica una medida de similitud distinta de la euclidiana, que en la práctica pondera la influencia de los vecinos en función de su proximidad al patrón que se requiere clasificar.

Esta, plantea (Molina & García, 2008) es una técnica de data mining basada en ejemplares en la que la medida de la distancia entre ejemplares se basa en la teoría de la información. Una forma intuitiva de verlo es que la distancia entre dos ejemplares se define como la complejidad de transformar un ejemplar en el otro. El cálculo de la complejidad se basa en primer lugar en definir un conjunto de transformaciones $T = t_1; t_2; \ldots; t_n$; σ para pasar de un ejemplo (valor de atributo) *a* a uno *b*.

La transformación es la de parada y es la transformación identidad ($\sigma(a) = a$). El conjunto *P* es el conjunto de todas las posibles secuencias de transformaciones descritos en T^* que terminan en σ y $\bar{t}(a)$ es una de estas secuencias concretas sobre el ejemplo *a*. Esta secuencia de transformaciones tendrá una probabilidad determinada $p(\bar{t})$, definiéndose la función de probabilidad $P^*(b|a)$ como la probabilidad de pasar del ejemplo *a* al ejemplo *b* a través de cualquier secuencia de transformaciones.

Luego de haber encontrado el clasificador se crearon los modelos multi etiquetas para crear el sistema de predicciones. Según (Spyromitros-Xioufis, Tsoumakas, Groves, & Vlahavas, 2012) el aprendizaje de etiquetas múltiples está estrechamente relacionado con la regresión de objetivos múltiples, también conocida como regresión multivariada o de múltiples salidas, tiene como objetivo predecir múltiples variables de destino con valores reales en lugar de variables binarias.

A pesar de que la regresión de múltiples objetivos es una tarea menos popular, todavía surge en varios dominios interesantes, como la predicción del ruido del viento de los componentes del vehículo, la predicción del precio de las acciones y el modelado ecológico. El aprendizaje de etiquetas múltiples a menudo se trata como un caso especial de regresión de objetivos múltiples en las estadísticas. Sin embargo, puede decirse que ambos son ejemplos de la tarea de aprendizaje más general de predecir múltiples objetivos, que pueden ser de valor real, binarios, ordinales, categóricos o incluso de tipo mixto. El enfoque de línea de base para aprender un modelo separado para cada objetivo se aplica a ambas tareas de aprendizaje.

Estos modelos fueron puestos a prueba, para ello se realizaron un total de 3 pruebas por modelos con el objetivo de verificar la efectividad en las predicciones de los modelos, los resultados de las pruebas se muestran en las tablas 4 y 5.

Tabla 4. Pruebas para *Leucaena leucocephala*.

Variable	Prueba 1		Prueba 2		Prueba 3	
	Real	**Esperado**	**Real**	**Esperado**	**Real**	**Esperado**
Tt	0,55	0,57	1,46	1,59	3,05	2,89
Tct	14,01	14,07	11,02	11,03	14,56	14,45
Tclt	11,15	11,19	9,33	9,32	10,79	10,74
Tcl	2,85	2,88	1,69	1,71	3,76	3,71
Ft	6,17	6,19	5,78	5,78	7,37	7,28

Vr	1,30	1,3	0,43	0,43	0,96	0,95
Es	0,50	0,5	0,18	0,18	0,46	0,46
Rf	2,03	2	1,21	1,21	1,19	1,19
Fl	11,80	11,81	28,61	28,73	37,67	37,86
Al	0,78	0,78	0,99	0,97	1,08	1,08
Sp	1,30	1,28	1,84	1,79	2,26	2,24
Tr	6,20	6,21	7,81	7,82	8,15	8,12
Et	7,14	7,2	11,81	11,72	13,63	13,54

Tabla 5. Pruebas para *Tithonia diversifolia*

Variable	Prueba 1		Prueba 2		Prueba 3	
	Real	Esperado	Real	Esperado	Real	Esperado
Tt	2,52	2,52	30,83	30,76	20,88	20,81
Tct	126,10	126,08	139,27	139,12	127,86	127,97
Tclt	117,17	117,15	130,35	130,22	119,21	119,32
Tcl	8,93	8,93	8,91	8,9	8,65	8,65
Ft	17,68	17,69	48,49	48,43	43,40	43,39
Vr	2,01	2,01	1,67	1,68	2,42	2,42
Es	2,08	2,079	3,66	3,66	3,05	3,05
Rf	2,27	2,22	1,79	1,79	1,80	1,8
Fl	30,41	30,44	61,12	61,14	77,86	77,85
Al	2,67	2,67	2,94	2,94	3,06	3,08
Sp	5,51	5,52	12,76	12,72	10,53	10,57
Tr	8,84	8,81	8,38	8,35	9,19	9,2
Et	5,39	5,38	5,29	5,22	8,52	8,71

Conclusiones

1. Con el estudio realizado se ha podido determinar una vía para estimar los componentes fitoquímicos de las variedades *Leucaena leucocephala* y *Tithonia diversifolia.* Si se observan los valores esperados y reales ser ve sin hacer mucho esfuerzo, que son

verdaderamente cercanos por lo que se puede definir con claridad que el modelo de predicciones es bastante acertado.

2. Al tener en cuenta que son muchos los factores que influyen en que la calidad de la planta sea óptima, se estima que mientras mejores sean las condiciones de clima, suelo y otros factores que influyen de forma directa en su salud, mejores serán los resultados finales en dependencia del fin que tengan.
3. La composición fitoquímica de las variedades de pastos y forrajes es determinante en la calidad de las mismas, estas no están exentas de la influencia de las condiciones del clima en su comportamiento. Es necesario recalcar que la calidad de los pastos y forrajes tropicales está condicionada por diferentes factores, entre los que se destacan especies, humedad, suelo, precipitaciones, temperatura y variedades entre otros.
4. Las plantas son fuentes de gran cantidad de productos metabólicos de importancia comercial y son usados en las industrias farmacéutica, alimenticia, de cosméticos y como fuentes de numerosas sustancias de interés agroquímico. Se estima que más de 100 000 metabolitos secundarios son producidos por las plantas. Aproximadamente 1 600 estructuras químicas nuevas obtenidas a partir de plantas superiores se describen cada año, de las cuales un gran número tiene actividad biológica. Esto hace que el estudio de los metabolitos presentes en las plantas sea un enorme reto, de ahí la necesidad de utilizar tecnologías diversas para su producción, caracterización e identificación (Pérez-Alonso & Jiménez, 2011).

Referencias Bibliográficas

Carmona Agudelo, J. C. (2007). Efecto de la utilización de arbóreas y arbustivas forrajeras sobre la dinámica digestiva en bovinos. *Lasallista de investigación*, *4*(1), 40–50.

Domínguez Gómez, T. G., Ramírez Lozano, R. G., Estrada Castillón, A. E., Scott Morales, L. M., González Rodríguez, H., & Alvarado, M. del S. (2012). Importancia nutrimental en plantas forrajeras del matorral espinoso tamaulipeco. *Ciencia UANL*, *15*(59), 77–93.

Figueroa, L. J. H., Manríquez, A. S., Melesio, J. A. G., Mendoza, J. G. H., & Pérez, N. V. R. (2018). Análisis de métodos de clasificación para el diagnóstico de fertilidad. *Pistas Educativas*, *36*(114).

Galindo, J., González, N., Marrero, Y., Sosa, A., Ruiz, T., Febles, G., … others. (2014). Efecto del follaje de plantas tropicales en el control de la producción de metano y la población de protozoos ruminales in vitro. *Revista Cubana de Ciencia Agrícola*, *48*(4).

García, D., Ojeda, F., & Montejo, I. (2003). Evaluación de los principales factores que influyen en la composición fitoquímica de Morus alba (Linn.). I Análisis cualitativo de metabolitos secundarios. *Pastos y Forrajes*, *26*(4).

Herrera, R., Verdecia, D., Ramírez, J., García, M., & Cruz, A. M. (2017). Relation between some climatic factors and the chemical composition of Tithonia diversifolia. *Revista Cubana de Ciencia Agrícola*, *51*(2), 271–279.

Molina, J., & García, J. (2008). Técnicas de Minería de Datos basadas en Aprendizaje Automático. *Técnicas de Análisis de Datos*, 96–266.

Pérez-Alonso, N., & Jiménez, E. (2011). Producción de metabolitos secundarios de plantas mediante el cultivo in vitro. *Biotecnología vegetal*, *11*(4).

Sepúlveda Jiménez, G., Porta Ducoing, H., & Rocha Sosa, M. (2003). La participación de los metabolitos secundarios en la defensa de las plantas. *Revista Mexicana de Fitopatología*, *21*(3).

Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W., & Vlahavas, I. (2012). Multi-label classification methods for multi-target regression. *arXiv preprint arXiv:1211.6581*, 1159–1168.

Vega Espinosa, M., Ramírez De la Ribera, J., Leonard Acosta, I., & Igarza, A. (2006). Rendimiento, caracterización química y digestibilidad del pasto Brachiaria decumbens en las actuales condiciones edafoclimáticas del Valle del Cauto. *Revista electrónica de Veterinaria REDVET*, *7*(5).