Modelos de predicción de metabolitossecundarios para dos variedades de plantas protéicas.

Secondary metabolite prediction models for two protein plant varieties.

Pedro Manuel Estrada Jiménez

Héctor Raúl González Diez

Alberto Verdecia Cabrera

Danis Manuel Verdecia Acosta

Jorge Luis Ramírez de la Rivera

Universidad de Granma. Cuba.

Universidad de las Ciencias Informáticas. La Habana. Cuba.

Universidad de Granma. Cuba.

Universidad de Granma. Cuba.

Universidad de Granma. Cuba.

Resumen

El presente trabajo se basa en la utilización de las librerías de Weka y mulan para determinar losmodelos de regresión ideales para un sistema de predicciones de metabolitos secundarios de dos variedades de plantas protéicas existentes en Cuba utilizadas en la alimentación animal y para los modelos de predicciones respectivamente. Los mismos fueron determinados haciendo uso de los clasificadores de weka en busca del modelo que retornara menor error cuadrático medio a partir de una base de casos.

Palabras clave: Clasificador, metabolito, predicción, protéica, regresión.

Abstract

The present work is based on the use of the libraries of Weka and mulan to determine the ideal regression models for a system of predictions of secondary metabolites for two varieties of protein plants existing in Cuba used in animal nutritionand for the models of predictions respectively. The same they were determined making use of weka's classifiers in search of the model that you return minor quadratic half an error as from a base of cases.

Keywords:Classifier,metabolite, prediction,protein,regression,

Introducción

En nuestro país la ganadería es uno de los principales sectores de la economía, siendo prioridad para el desarrollo nacional. La exportación de sus derivados es utilizada en el turismo y otras áreas. Es válido aclarar que no solo la producción vacuna se beneficia de las investigaciones desarrolladas en el ámbito de la nutrición animal: la porcina, equina y ovina también se favorecen con el desarrollo de proyectos de innovación e investigación.

Universidades cubanas, en colaboración con diferentes países e instituciones, realizan proyectos encaminados a la investigación de los nutrientes de las plantaciones que actualmente son utilizadas en la alimentación animal. En la provincia Granma existen extensiones destinadas al cultivo de distintas variedades de plantas protéicas como la localizada en el Valle del Cauto, donde existe una población de las variedades Leucaenaleucophala y Tithoniadiversofilia.

Los metabolitos secundarios son compuestos químicos sintetizados que cumplen funciones no esenciales en las plantas. Estos son un mecanismo de defensa ante la posibilidad de que estas sean usadas en la alimentación animal. El comportamiento de estos en plantas protéicas es una herramienta teniendo en cuenta los efectos que pueden ocasionar en caso de una mala administración, las plantas que contienen un determinado nivel de algunos de estos pueden causar efectos secundarios en animales (García, 2004), (Sepúlveda Jiménez, Porta Ducoing, & Rocha Sosa, 2003). Existe una gran cantidad de leguminosas arbóreas y arbustivas tropicales que contienen factores o componentes antinutricionales como los taninos, saponinas y otros metabolitos secundarios que dificultan la digestibilidad en los animales, en especial los taninos (Carmona Agudelo, 2007).Estudios realizados han arrojado valores positivos como los descritos por (García, Ojeda, & Montejo, 2003) donde se caracteriza la composición fitoquímica en la fracción comestible de Tithoniadiversofilia.

Teniendo en cuenta la importancia de las plantaciones para la alimentación ganadera, el Centro de Estudios de Producción Animal de la Universidad de Granma ha desarrollado investigaciones encaminadas a conocer con mayor exactitud los metabolitos secundarios de las variedades antes mencionadas. Este trabajo tiene como objetivo buscar, partiendo de una base de conocimientos facilitada por los científicos del Centro de Estudios de Producción Animal, un modelo de regresión para un sistema de predicciones de los metabolitos secundarios en el cual se genere el menorerror cuadrático medio posible. Es necesario

destacar que estos estudios de determinación de metabolitos secundarios son muy costosos en tiempo y económicamente, además. Para realizarlos se necesita de equipos de difícil adquisición.

Entre los estudios más recientes realizados se puede citar el trabajo de (Verdecia Acosta et al., 2014) donde utiliza el análisis multivariado de conglomerados para agrupar especies con características similares dentro de las que se encuentran las variedades estudiasen este artículo.

Materiales y métodos

Para el estudio se tuvieron en cuenta una serie de variables de entrada (variables independientes) y salida (variables dependientes o metabolitos secundarios) que se codificaron como se muestra en la tabla 1. En este punto los valores de cada una de las variables de entrada tienen rangos diferentes, por ejemplo, la edad, oscila entre 60 y 180 y las temperaturas entre 10 y 40. Es válido aclarar que el período es una variable que se refiere a la etapa del año en que se realiza el experimento. Esto significa que el mismo se realizó en los períodos del año que son lluvia y poca lluvia, los cuales fueron codificados, determinándose para el caso de período de lluvia 0 y para poca lluvia 1; no obstante, estos no dejan de ser valores reales, por tanto, todos los valores del estudio realizado pertenecen al dominio de los números reales donde se puede decir que para cualquier valor de las variables de entrada o salida $x \epsilon \mathbb{R} / x > 0$. Esto podrá verse a partir de las tablas de las pruebas aplicadas a los modelos.

Tabla 1. Relación de variables creadas para cada uno de los parámetros a evaluar.

	Variable	Código	Dominio y Rango
Variables independientes	Período	periodo	$x \epsilon \mathbb{R} / x = 0 \text{ ó } x = 1$
	Edad	edad	$\mathbb{R}^+$
	Nitrógeno	nitrogeno	$\mathbb{R}^+$
	Glucosa	glucosa	$\mathbb{R}^+$
	Fructosa	fructosa	$\mathbb{R}^+$
	Sacarosa	sacarosa	$\mathbb{R}^+$
	TemperaturaMáxima	temperaturaMax	$\mathbb{R}^+$
	TemperaturaMínima	temperaturaMin	$\mathbb{R}^+$
	TemperaturaMedia	temperaturaMed	$\mathbb{R}^+$
	HumedadRelativaMáxima	humedadRelativaMax	$\mathbb{R}^+$
	HumedadRelativa Mínima	humedadRelativaMin	$\mathbb{R}^+$
	HumedadRelativaMedia	humedadRelativaMed	$\mathbb{R}^+$
	Lluvia	lluvia	$\mathbb{R}^+$
	Días conLluvia	diasConLluvia	$\mathbb{R}^+$

Variables dependientes	TaninosTotales	taninosTotales	$\mathbb{R}^+$
	TaninosCondensadosTotales	taninosCondensadosTotales	$\mathbb{R}^+$
	TaninosCondensadosLigadosTotales	taninosCondensadosLigadosTotales	$\mathbb{R}^+$
	TaninosCondensadosLibres	taninosCondensadosLibres	$\mathbb{R}^+$
	FenolesTotales	fenolesTotales	$\mathbb{R}^+$
	Verbascosa	verbascosa	$\mathbb{R}^+$
	Estaquiosa	estaquiosa	$\mathbb{R}^+$
	Rafinosa	rafinosa	$\mathbb{R}^+$
	Flavonoides	flavonoides	$\mathbb{R}^+$
	Alcaloides	alcaloides	$\mathbb{R}^+$
	Saponinas	saponinas	$\mathbb{R}^+$
	Triterpenos	triterpenos	$\mathbb{R}^+$
	Esteroides	esteroides	$\mathbb{R}^+$

Los valores referentes a la investigación fueron cedidos por el personal científico que labora en el Centro de Estudios de Producción Animal de la Universidad de Granma. Los mismos realizaron varios experimentos en base a distintos períodos de edades de rebote que consiste en realizar un corte a una de las plantas estudiadas en este caso y determinar, con lo extraído de ellas, los metabolitos secundarios a partir de reactivos y otros recursos y materiales químicos con los que no se cuenta en la Universidad.

Por tal motivo, surge la necesidad de crear un sistema que con los valores proporcionados por los investigadores sea capaz de predecir los metabolitos secundarios. Para ello se estudiaron varias soluciones entre las que destacan varios modelos de regresión lineal de tal forma que fueran directamente proporcional a la cantidad de variables de salida, uso de redes neuronales artificiales y modelos de regresión multi target donde se estudiaron los algoritmos de clasificación y vías planteados en (Spyromitros-Xioufis, Tsoumakas, Groves, & Vlahavas, 2012). Con este juego de datos se establece un modelo multilabel donde se conocen todas las etiquetas (variables mencionadas anteriormente) y contiene un conjunto de datos finitos que forman el conjunto de datos para el aprendizaje como se plantea en (Tsoumakas, Spyromitros-Xioufis, Vrekou, & Vlahavas, 2014). Para todos se plantearon soluciones varias, pero el resultado final estaba encaminado a buscar un modelo que proporcionara el menor error cuadrático medio para con este crear el modelo de predicciones. Entre todos los estudios analizados el que mejores resultados arrojó fue el uso del modelo multi target. Para ello se utilizaron clasificadores de weka basándose en algunos de los ejemplos planteados en (Tsoumakas, Spyromitros-Xioufis, Vrekou, & Vlahavas, 2014) y (Read, Reutemann, Pfahringer, & Holmes, 2016). Para el proceso de selección de los clasificadores a utilizar, se cargaron los datos de entrenamiento del modelo en weka y de ahí se procedió a buscar los clasificadores de weka que estuvieran habilitados para este juego de datos.El clasificador más eficiente a partir del RMSE fue lazy.KStar que tuvo 0,0933±0 y 0,0783±0 para Leucaenaleucophala y Tithoniadiversofilia respectivamente.

Resultados y discusión

Luego de haber encontrado un modelo ideal para las variedades que se estudian en cuestión se ejecuta el paso de hacer pruebas para ver cuán acertadas son las predicciones. En este paso se utilizó Multi-Target Stacking de mulan, se analizaron previamente las documentaciones de los modelos Multi Target. Luego se ejecutaron un total de seis pruebas por modelos para evaluar las diferencias entre los valores arrojados por la predicción (valor real) y los valores que en sí debería dar el sistema una vez ejecutada la predicción (valor esperado). Los resultados obtenidos de los test aplicados se muestran en las tablas 2, 3, 4 y 5.

Tabla 2. Pruebas del modelo de regresión para Leucaena leucophala.

Variable	Prueba 1		Prueba 2		Prueba 3	
	Real	Esperado	Real	Esperado	Real	Esperado
taninosTotales	0,5522040146856	0,57	5,3515108920097	5,42	3,0457847935339	2,89
taninosCondensadosTotales	14,0087419011318	14,07	13,6056738783941	13,71	14,5567263837650	14,45
taninosCondensadosLigadosTotales	11,1522099156848	11,19	9,5167122770833	9,57	10,7913943395373	10,74
taninosCondensadosLibres	2,8548764485692	2,88	4,0893234550351	4,14	3,7645802342942	3,71
fenolesTotales	6,1694991560646	6,19	12,3665343177053	12,53	7,3650707190360	7,28
verbascosa	1,3015848983242	1,3	1,1118638120664	1,12	0,9610789395993	0,95
estaquiosa	0,5007190383707	0,5	0,3183085275897	0,32	0,4592795968250	0,46
rafinosa	2,0300251788802	2	0,9595494878760	0,97	1,1879803334019	1,19
flavonoides	11,7966606817078	11,81	24,6590897394088	24,72	37,6729705013467	37,86
alcaloides	0,7809058860416	0,78	0,9019647654323	0,9	1,0787945991336	1,08
saponinas	1,2983553256239	1,28	1,3384750312975	1,39	2,2572636161791	2,24
triterpenos	6,1966436764719	6,21	7,6254526349485	7,65	8,1451586341549	8,12
esteroides	7,1439690296216	7,2	10,6254849028744	10,57	13,6272515047000	13,54

Resultados referentes al período de Lluvia

Resultados referentes al período de Lluvia

Tabla 3. Pruebas del modelo de regresión para Leucaenaleucophala.

Variable	Prueba 4		Prueba 5		Prueba 6	
	Real	Esperado	Real	Esperado	Real	Esperado
taninosTotales	2,1759082709239	2,18	1,4571243862794	1,59	3,0344187685810	3,04
taninosCondensadosTotales	10,3751248252045	10,38	11,0218461544002	11,03	13,0458499739440	13,04
taninosCondensadosLigadosTotales	8,7267710909181	8,73	9,3276688824232	9,32	10,1039068197650	10,09
taninosCondensadosLibres	1,6515528845402	1,65	1,6943277101565	1,71	2,9442811856749	2,95
fenolesTotales	5,3982959712347	5,38	5,7823776573976	5,78	6,4862795025470	6,45
verbascosa	0,7008390166917	0,7	0,4312237679208	0,43	0,3179791437385	0,32
estaquiosa	0,2102152113918	0,21	0,1843495102207	0,18	0,2131537012613	0,21
rafinosa	1,6008365182747	1,6	1,2101781095600	1,21	1,4682626948564	1,47
flavonoides	15,2991797584860	15,29	28,6123655778639	28,73	44,0318136924470	43,96
alcaloides	0,7818464942665	0,79	0,9944081644208	0,97	1,1528028308403	1,17
saponinas	1,6878339811939	1,67	1,8354735247763	1,79	2,3658627817016	2,35
triterpenos	6,2640581840928	6,27	7,8072338187442	7,82	9,1781643777703	9,2
esteroides	8,6622323291772	8,68	11,8091730190183	11,72	14,8454171470492	14,83

Resultados referentes al período de Poca Lluvia

Tabla 4. Pruebas del modelo de regresión para Tithoniadiversofilia.

Variable	Prueba 1		Prueba 2		Prueba 3	
	Real	Esperado	Real	Esperado	Real	Esperado
taninosTotales	2,5235203509512	2,52	22,4236307888475	22,35	20,8811579654434	20,81
taninosCondensadosTotales	126,1032983393340	126,08	132,0936007142940	132,1	127,8593912910620	127,97
taninosCondensadosLigadosTotales	117,1732876109660	117,15	121,6095316883620	121,6	119,2126234133520	119,32

taninos-Condensados-Libres	8,9300372062014	8,93	10,4841066508967	10,5	8,6467215650382	8,65
fenoles-Totales	17,6790859811669	17,69	44,0077471395816	44,01	43,4018183223873	43,39
verbascosa	2,0115773281604	2,01	4,3531390909053	4,36	2,4237007265365	2,42
estaquiosa	2,0802316432289	2,079	4,4132751597171	4,41	3,0500172007812	3,05
rafinosa	2,2661638586671	2,22	2,2007912127863	2,08	1,8031905968155	1,8
flavonoides	30,4149992208030	30,44	59,2096657628074	59,25	77,8571124543536	77,85
alcaloides	2,6699340710007	2,67	2,8801418977883	2,86	3,0598732573710	3,08
saponinas	5,5086563161363	5,52	8,6180392313229	8,89	10,5280959800069	10,57
triterpenos	8,8375257003432	8,81	7,7459171423255	7,79	9,1907171085036	9,2
esteroides	5,3893343386896	5,38	5,9306034346766	5,91	8,5205964275754	8,71

Resultados referentes al período de Lluvia

Tabla 5. Pruebas del modelo de regresión para Tithoniadiversofilia.

Variable	Prueba 4		Prueba 5		Prueba 6	
	Real	Esperado	Real	Esperado	Real	Esperado
taninosTotales	23,5841167526855	23,51	30,8262871641865	30,76	34,0608967906895	34,02
taninosCondensadosTotales	127,9488609814830	128,28	139,2687482867760	139,12	142,6552236764930	142,53
taninosCondensadosLigadosTotales	117,5652140318990	117,93	130,3540832055570	130,22	131,2581274231940	131,18

taninosCondensadosLibres	10,3883264334332	10,35	8,9143091742229	8,9	11,3956878743356	11,35
fenolesTotales	44,3468593093546	44,3	48,4940853318984	48,43	50,4753447428539	50,46
verbascosa	2,9474165393461	2,95	1,6706398115015	1,68	1,1246335268727	1,12
estaquiosa	3,5686950947866	3,574	3,6558232241312	3,66	0,1813547704044	0,18
rafinosa	1,8900244878655	1,89	1,7945429165827	1,79	0,9785847826765	0,98
flavonoides	46,9783919586474	46,97	61,1184103222655	61,14	86,9840761057273	86,97
alcaloides	2,7868929536194	2,79	2,9445792516338	2,94	3,2497080526290	3,25
saponinas	7,7633829902147	7,74	12,7638776664054	12,72	15,4080982855569	15,36
triterpenos	7,1099989248352	7,05	8,3803866050747	8,35	9,1372753550800	9,17
esteroides	3,1401659558689	3,12	5,2931256007662	5,22	8,4822492261519	8,52

Resultados referentes al período de Poca Lluvia

Si valoramos los valores esperados y reales nos damos cuenta sin hacer mucho esfuerzo de que son verdaderamente cercanos por lo que se puede definir con claridad que el modelo de predicciones es bastante acertado. Los especialistas del departamento de Pastos y Forrajes de la Universidad de Granma corroboraron los valores obtenidos en las predicciones en base a la exactitud de las mismas y la similitud con los valores esperados.

Conclusiones

Con el desarrollo de este trabajo se pudieron determinar los modelos de predicciones de metabolitos secundarios adecuados o ideales para dos variedades de plantas proteicas, los mismos partieron de bases de datos almacenadas en un período de dos años donde se guardaban valores referentes a los metabolitos (variables dependientes) y las variables a partir de las cuales se debía extraer la información deseada (variables independientes), datos que sirvieron para el aprendizaje y validación de los modelos. El estudio presentado ayuda a tener un mayor control de estos en las plantas protéicas estudiadas para la alimentación animal; la utilización de los modelos encontrados servirá para tener, más allá de predecir valores, una idea del comportamiento de estos (metabolitos secundarios).

Agradecimientos

A Dios por ayudarme a ser quien soy, a mis padres por haberme traído al mundo e inculcarme la educación como fuente de progreso. A mi esposa e hijas por ser mi motor impulsor y mi razón de ser en la vida. Agradezco a la Revolución por haberme dado la posibilidad de formarme como Ingeniero en Ciencias Informáticas en la Universidad de las Ciencias Informáticas y al Comandante en Jefe Fidel Castro Ruz por haber creado tan bella e importante casa de altos estudios. A mis tutores DrC. Yolanda Soler Pellicer, DrC. Danis Manuel Verdecia Acosta, Jorge Luis Ramírez de la Rivera, a todos mis profesores de la Maestría en Ciencia de la Computación de la Universidad de Oriente, mis inolvidables compañeros de cuarto, Ing. José Antonio Leyva Regalón, Ing. Henrry Tomás Brown Grandales y Lic. Asdruval Henry Nelson, mi hermano y toda mi familia por su apoyo incluyendo a Erodis Pérez Michel. A todas las personas que de alguna forma han tenido que ver con la realización de este trabajo.

Referencias

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

García, D. (2004). Los metabolitos secundarios de las especies vegetales. Pastos y forrajes, 27(1).

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).

Read, J., Reutemann, P., Pfahringer, B., & Holmes, G. (2016). Meka: a multi-label/multi-target extension to weka. The Journal of Machine Learning Research, 17(1), 667–671.

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 e-prints.

Tsoumakas, G., Spyromitros-Xioufis, E., Vrekou, A., & Vlahavas, I. (2014). Multi-target regression via random linear target combinations. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 225–240).

Verdecia Acosta, D. M., Herrera García, R. S., Ramírez de la Ribera, J. L., Acosta, I. L., Bodas Rodríguez, R., Lorente, S. A., … others. (2014). Caracterización bromatológica de seis especies forrajeras en el Valle del Cauto, Cuba. Avances en Investigación Agropecuaria, 18(3).