Investigación

Aplicación de la regresión de múltiples objetivos en la estimación de componentes fitoquímicos

Application of multi-target regression in the estimation of phytochemical components

Pedro Manuel Estrada Jiménez
Pedro Jorge Noguera López
Raúl Recio Avilés

Revista de Investigación

Volumen X, Número 2, pp. 007–014, ISSN 2174-0410
Recepción: 01 May'20; Aceptación: 25 May'20

1 de octubre de 2020

Resumen

La aplicación de los modelos de regresión es un paradigma que está en constante evolución en gran parte de los países. En la presente investigación se hace uso de los modelos de regresión de múltiples objetivos en la predicción de los componentes fitoquímicos de dos variedades de plantas utilizadas en la nutrición animal. Para esta variante se utilizó un algoritmo basado en instancias. Se evaluó el aprendizaje de este modelo para determinar la calidad de las predicciones y finalmente se obtuvieron los modelos de predicciones para ambas variedades.

Palabras Clave: Múltiples objetivos, regresión, aprendizaje automático, minería de datos.

Abstract

The application of regression models is a paradigm that is constantly evolving in many countries. In the present research, multi-objective regression models are used in the prediction of the phytochemical components of two varieties of plants used in animal nutrition. For this variant, an instance-based algorithm was used. The learning of this model was evaluated to determine the quality of the predictions and finally the prediction models for both varieties were obtained.

Keywords: Multi-target, regression, machine learning, data mining.

1. Introducción

El sector agropecuario a nivel mundial es uno de los más importantes en el desarrollo de los países, por lo que ocupa uno de los primeros lugares en los renglones a valorar; las tecnologías

actuales se han dedicado a la fertilización y obtención de plantaciones a partir de la aplicación de métodos modernos que han provocado en gran medida la pérdida de cualidades necesarias y naturales de algunos cultivos, sin embargo las propiedades de los árboles y arbustos contribuyen a asegurar una dieta nutritiva para el ganado [3].

En investigaciones realizadas ha quedado reflejada la importancia del uso de estas plantas en sistemas silvopastoriles mediante la aplicación de varias técnicas y la importancia del control de los componentes fitoquímicos que estas contienen ante la aparición de efectos secundarios en animales atendiendo al consumo de plantas denominadas de excelencia. Estas han generado en gran cantidad de animales trastornos digestivos con distintas manifestaciones producto del efecto de los metabolitos secundarios.

La estimación de estos componentes es un proceso costoso, en la actualidad el mismo se realiza mediante técnicas de laboratorio caras. El uso de sistemas inteligentes ha tenido su participación en la toma de decisiones e investigaciones en varios ámbitos, se pueden mencionar resultados vinculados con la medicina, electrónica, meteorología y otros; la agricultura y la ganadería no escapan de estos beneficios. La aplicación de las nuevas tecnologías y de la Inteligencia Artificial en la solución de problemas genera en el mundo gran impacto en los sectores donde se aplican las soluciones por la rapidez y la eficiencia con que se desarrollan las tareas.

Entre los avances en la aplicación de los modelos de regresión se destacan a nivel mundial los modelos de predicciones en todas sus variantes, estos, se aplican a un gran número de problemas. En este sentido los clasificadores juegan un papel importante en varias esferas como en la determinación de patologías en el ámbito de la medicina y la agricultura donde cada día son más explotados y estudiados. Los algoritmos de clasificación se pueden dividir en dos grupos, supervisados y no supervisados. Las técnicas supervisadas tienen una fase de entrenamiento en la cual se usan muestras representativas de las clases seleccionadas para establecer un modelo del proceso de clasificación. Las técnicas no supervisadas no requieren ningún entrenamiento y tampoco suponen la definición previa de una clase.

El impacto de la aplicación de técnicas científicas vinculadas con los sectores de la sociedad forma parte del desarrollo del país; las predicciones han alcanzado un peldaño en la cima de estas; se puede mencionar la aplicación de los modelos de regresión en la estimación del estado del tiempo y las variables climáticas. Esta tarea se realiza auxiliándose de un historial de pronósticos dotado de un gran volumen de datos que permite a los sistemas inteligentes entrenarse y alcanzar un conocimiento alto para poder predecir el estado del tiempo.

2. Materiales y métodos o Metodología computacional

En el proceso de selección de la metodología a utilizar fueron exploradas la naturaleza de las variables a estudiar, de esto se derivó que las mismas estaban todas en el dominio de los números reales por lo que se concluyó que todas las variables, tanto las de entrada como las de salida estaban en el mismo dominio numérico. Partiendo de este paso se evaluó la correlación entre estas con el objetivo de determinar el posible modelo de regresión a aplicar; de este se determinó que como la correlación entre las variables de salida era alta entonces es recomendable la utilización de un modelo de regresión de múltiples objetivos atendiendo a la correlación entre las variables y el comportamiento de las mismas. Estos modelos de regresión son utilizados en la predicción como técnica supervisada en la minería de datos en la cual se emplean algoritmos que resuelvan problemas de regresión.

Los modelos de regresión donde se tiene varias salidas hacen uso de un algoritmo para realizar las predicciones de cada objetivo. Dentro de estos algoritmos pueden encontrarse:

Los **algoritmos basados en reglas** permiten expresar disyunciones de manera más fácil que los árboles y tienden a preferirse con respecto a los árboles por tender a representar partes

de conocimiento relativamente independientes. Las técnicas de Inducción de Reglas permiten generar y contrastar árboles de decisión, o reglas y patrones a partir de los datos de entrada. La información de entrada será un conjunto de casos en que se ha asociado una clasificación o evaluación a un conjunto de variables o atributos [1].

Los **algoritmos basados en árboles de decisión** son un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que permite determinar la decisión final que se debe tomar al seguir las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas. Los árboles de decisión son especialmente apropiados para expresar procedimientos médicos, legales, comerciales, estratégicos, matemáticos, lógicos, entre otros. Estos se caracterizan por la sencillez de su representación y de su forma de actuar, además de la fácil interpretación, dado que pueden ser expresados en forma de reglas de decisión [1].

Los **algoritmos perezosos** son métodos basados en instancias que utilizan enfoques conceptualmente sencillos para las aproximaciones de valores reales o discretos de las funciones de salida. Aprender en estos modelos consiste en almacenar los datos de entrenamiento presentados y cuando una nueva instancia es encontrada, un grupo de ejemplos similares relacionados son recuperados de memoria y usados para clasificar la nueva instancia consultada. Entre los algoritmos perezosos destacan los modelos de aprendizaje basados en instancias. Su funcionamiento parte de almacenar instancias de ejemplo, que en algunas variantes son todas las instancias del conjunto de entrenamiento, en otras solo se almacenan los ejemplares más representativos, etc. [7].

En este sentido en el algoritmo de aprendizaje basado en instancias, el funcionamiento es muy simple: se almacenan los ejemplos de entrenamiento de datos históricos y cuando se requiere clasificar a un nuevo objeto, se extraen los objetos más parecidos y se usa su clasificación para clasificar al nuevo objeto. Los vecinos más cercanos a una instancia se obtienen en dependencia de los atributos, para el caso de valores continuos se utiliza la distancia Euclidiana sobre los n posibles atributos y el resultado de la clasificación puede ser discreto o continuo; en el caso discreto, el resultado de la clasificación es la clase más común de los k vecinos [5].

3. Regresión de múltiples objetivos

Existen varias clasificaciones de los modelos de regresión, esto está condicionado por la naturaleza de las variables que intervengan en el problema y por la cantidad y organización de las mismas. Se puede decir que cuando se tiene una variable dependiente y una independiente puede aplicarse un modelo de regresión lineal simple, teniendo en cuenta los supuestos que normalmente se consideran y estudian para que esto se cumpla, de manera similar cuando se tienen varias variables de entrada y una de salida o varias variables independientes y una dependiente se puede plantear un modelo de regresión múltiple univariado pero cuando se tienen varias variables de entrada y varias de salida entran en juego otras técnicas a analizar para poder determinar la variante a aplicar, en dependencia del comportamiento de la correlación entre los datos es que se define la variante de regresión, pues cuando las variables de salida tienen una alta correlación entre ellas se puede deducir un problema de regresión de múltiples objetivos pero si la correlación es baja entonces se puede pensar en crear tantos modelos múltiples univariados como variables de salida se tengan, pero siempre hay que tener en cuenta la correlación que exista entre las variables que intervienen en el problema planteado.

Normalmente un problema de regresión de múltiples objetivos está compuesto por un conjunto de datos S que contiene todos los ejemplos en la forma (x, y) donde $x \in X$ es un vector de entrada y $y \in Y$ es uno destino. X es el espacio de entrada que contiene d variables de entrada $\{X_1, X_2, .., X_d\}$ y Y es el espacio de salida que contiene q variables objetivo $\{Y_1, Y_2, .., Y_q\}$ Por tanto x_i es el vector de entrada del ejemplo i y y_i representa el vector objetivo del ejemplo i . Por tanto, dados los ejemplos de entrenamiento $S = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ con d ejemplos de

entrenamiento, el objetivo de estos modelos es aprender de un modelo predictivo que dado un vector de entrada x sea capaz de predecir un vector objetivo $\hat{y}$ que se aproxime lo mejor posible al vector del cual aprendió el sistema [6].

Los métodos existentes para la regresión de múltiples salidas se pueden clasificar como métodos de transformación de problema (también conocidos como métodos locales), que transforman la salida múltiple del problema en problemas independientes de salida única, cada uno resuelto usando una salida única del algoritmo de regresión, y métodos de adaptación de algoritmos (también conocidos como métodos globales o big-bang) que adaptan un método específico de salida única (como los árboles de decisión y soporte de máquinas de vectores) para manejar directamente conjuntos de datos de múltiples salidas. Se considera que los métodos de adaptación son más desafiantes ya que generalmente no apuntan solo para predecir los objetivos múltiples, sino también para modelar e interpretar las dependencias entre estos objetivos [2].

Los métodos de transformación de problemas se basan principalmente en transformar el problema de regresión de múltiples salidas en problemas de un solo objetivo, luego construir un modelo para cada objetivo y finalmente concatenar todas las predicciones. El principal inconveniente de estos métodos es que se ignoran las relaciones entre las salidas y se predicen estas de manera independiente, situación que puede afectar a la calidad general de las predicciones. Entre los métodos de regresión de múltiples salidas se pueden mencionar:

- Método de un solo objetivo.
- Método de apilamiento de regresores de múltiples objetivos.
- Método de cadenas de regresores.
- Método de regresión de vectores de soporte de múltiples salidas.

Los métodos de adaptación de algoritmos se basan en la idea de predecir varias salidas de manera simultánea utilizando un modelo simple pero explorando las dependencias entre cada una de ellas (las salidas). Entre estos métodos se pueden mencionar:

- Método estadístico.
- Método de cadenas de regresores.
- Método de núcleos.
- Método basado en árboles de regresión.
- Método basado en reglas de clasificación.

De acuerdo con [9] para evaluar los modelos de regresión donde se tiene múltiples salidas la métrica más utilizada es RRMSE (Relative Root Mean Squared Error) . Esto se define como la raíz cuadrada de la distancia cuadrada promedio entre el puntaje real y el puntaje predicho:

$$RRMSE = (h, D_{prueba}) = \sqrt{\frac{\sum_{(x,y)\in D_{prueba}} (\hat{y}_j - y_j)^2}{\sum_{(x,y)\in D_{prueba}} (\bar{Y}_j - y_j)^2}}$$

donde $\bar{Y}_j$ es el valor medio de la variable de destino Y_j sobre $D_{entrenamiento}$ y $\hat{y}_j$ es la estimación de $h(x)$ para Y_j, $D_{entrenamiento}$ es el conjunto de entrenamiento y D_{prueba} el conjunto de prueba. El cálculo de RRMSE para un objetivo es igual al Error Cuadrático Medio (Root Mean

Squared Error RMSE) para ese objetivo dividido por el RMSE de predecir el valor promedio de ese objetivo en el conjunto de entrenamiento. La medida RRMSE se estima al utilizar el enfoque de espera para conjuntos de datos grandes, mientras que la validación cruzada de 10 veces se emplea para conjuntos de datos pequeños.

A menudo estos métodos de aprendizaje automático son aplicados con la validación cruzada, que es un procedimiento de remuestreo que se utiliza para evaluar modelos de aprendizaje automático en una muestra de datos limitada. El procedimiento tiene un único parámetro llamado k que se refiere al número de grupos en que se dividirá una muestra de datos dada. Este es un método que intenta maximizar el uso de los datos disponibles para la capacitación y luego probar un modelo. Es particularmente útil para evaluar el rendimiento del modelo, ya que proporciona un rango de puntajes de precisión a través de conjuntos de datos diferentes [4]. Por tanto se tiene entonces que para calcular el aRRMSE (Average Root Mean Squared Error) se promedia de la siguiente forma [8]

$$aRRMSE(h, D_{prueba}) = \frac{1}{q} \sum_{j=1}^{q} RRMSE$$

Esta medida de evaluación del aprendizaje es muy utilizada para evaluar el aprendizaje de los modelos de múltiples salidas, cuando el conjunto de datos es pequeño se utiliza la técnica de validación cruzada con 10 partes.

4. Resultados y discusión

A partir de un estudio realizado en el cual se tuvo en cuenta la necesidad de determinar los metabolitos secundarios, fueron analizados los datos necesarios para obtener los mismos por lo que se definieron como variables de entrada a las variables $E, N, Gl, Fr, Sc, TMax, TMin, TMed, HRMax, HRMin, HRMed, LlyDLl$ (explicadas en la Tabla 1) variables de salida, $Tt, Tct, Tclt, Tcl, Ft, Vr, Es, Rf, Fl, Al, Sp, Tr$ y Et (explicadas en la Tabla 2).Luego se conformaron los modelos para cada dataset (uno para el caso de la *Leucaena leucocephala* y otro para el caso de *Tithonia diversifolia*) a partir de un modelo de regresión de múltiples objetivos. El aprendizaje de estos fue probado utilizando algoritmos que modelan problemas de regresión como puede verse en la Tabla 3. Esta evaluación se realizó a partir de la aplicación de la validación cruzada con kfold=10 donde la medida utilizada para estos modelos es el aRRMSE.

Tabla 1. Variables de entrada.

Componente	Dominio $x \in \mathbb{R}$
Edad de rebote (E)	$\mathbb{R}+$
Nitrógeno (N)	$\mathbb{R}+$
Glucosa (Gl)	$\mathbb{R}+$
Fructuosa (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 (Ll)	$\mathbb{R}+$
Días con Lluvia (DLl)	$\mathbb{R}+$

Tabla 2. Variables de salida.

Componente	Dominio $y \in \mathbb{R}$
Taninos Totales (Tt)	$\mathbb{R}$
Taninos Condensados Totales (Tct)	$\mathbb{R}$
Taninos Condensados Ligados Totales (Tclt)	$\mathbb{R}$
Taninos Condensados Libres (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)	$\mathbb{R}$

Tabla 3. Clasificadores para Leucaena leucocephala y Tithonia diversifolia.

Algoritmo	RMSE para *Leucaena leucocephala*	RMSE para *Tithonia diversifolia*
M5p	0,7477	0,1789
m5pRuler	0,462	0,1243
linearRegression	1,069	0,1466
Ibk	**0,1176**	**0,0814**
ZeroR	4,6358	0,8386
Ksvm	0,2505	0,1102
SMOreg	1,5003	0,1481
DecisionStump	2,8875	0,9383
MultilayerPerceptron	0,1505	0,0945
GaussianProcesses	1,6809	0,2693
REPTree	0,4527	0,1469

Tabla 4. Resumen para Leucaena leucocephala y Tithonia diversifolia.

Métrica	Promedios para *Leucaena leucocephala*	Promedios para *Tithonia diversifolia*
Average RMSE	0,1176	0,0814
Average Relative RMSE	0,0595	0,0956
Average MAE	0,0814	0,0547
Average Relative MAE	0,0485	0,0697

Como puede observarse en la Tabla 3, el comportamiento del aprendizaje a partir del RMSE varía en dependencia del tipo de algoritmo con que fue probado. El menor valor obtenido fue con el IBK; este utiliza la técnica del vecino más cercano (KNN), es un algoritmo simple que almacena todos los casos disponibles y predice el objetivo numérico en función de una medida de similitud (por ejemplo, funciones de distancia). Los valores de la Tabla 4 muestran las principales medidas de desempeño para un modelo de regresión.

Para cada modelo se realizaron un total de 3 pruebas, la tabla que a continuación se muestra representa las pruebas realizadas con los modelos creados. Para este caso de prueba el objetivo principal se centró en la comprobación de la exactitud de las predicciones a partir de juegos de datos con los que el sistema no fue entrenado.

En consulta con especialistas en el área de Pastos y Forrajes de la Universidad de Granma se pudo comprobar que los resultados arrojados por las pruebas a los modelos se acercan bastante

Tabla 5. Resumen para Leucaena leucocephala y Tithonia diversifolia.

	Leucaena leucocephala				*Tithonia diversifolia*			
	Lluvia		Poca Lluvia		Lluvia		Poca Lluvia	
	Esp	Real	Esp	Real	Esp	Real	Esp	Real
Tt	22,35	22,34	30,76	30,78	0,57	0.55	1,59	1.47
Tct	132,1	132,18	139,12	139,24	14,07	14.01	11,03	11.0
Tclt	121,6	121,61	130,22	130,33	11,19	11.16	9,32	9.31
Tclt	10,5	10,52	8,9	8,91	2,88	2.85	1,71	1.69
Ft	44,01	44,03	48,43	48,45	6,19	6.15	5,78	5.76
Vr	4,36	4,36	1,68	1,67	1,3	1.3	0,43	0.44
Es	4,41	4,42	3,66	3,66	0,5	0.5	0,18	0.18
Rf	2,08	2,1	1,79	1,79	2	2.03	1,21	1.21
Fl	59,25	59,21	61,14	61,11	11,81	11.79	28,73	28.83
Al	2,86	2,88	2,94	2,95	0,78	0.78	0,97	0.99
Sp	8,89	8,7	12,72	12,75	1,28	1.3	1,79	1.8
Tr	7,79	7,76	8,35	8,38	6,21	6.19	7,82	7.82
Et	5,91	5,91	5,22	5,27	7,2	7.15	11,72	11.8

a los arrojados por las técnicas de laboratorio. Este análisis fue posible a partir de las pruebas mostradas en la Tabla 5.

5. Conclusiones

- Se crearon modelos de regresión de múltiples objetivos capaces de predecir los componentes fitoquímicos de las especies *Leucaena leucocephala* y *Tithonia diversifolia* a partir de factores climáticos, metabolitos primarios y edad de rebote, en las condiciones climáticas del Valle del Cauto, esto posibilitará a científicos y a todo aquel que desee conocer el comportamiento de las especies mencionadas agilizar el proceso de caracterización fitoquímica de estas especies y sustituirá el uso de técnicas de laboratorio costosas para esta tarea.
- Se evaluó el aprendizaje de los modelos mediante la validación cruzada para obtener el modelo con menor aRRMSE.

Referencias

[1] Yadira Robles Aranda and Anthony R Sotolongo, *Integración de los algoritmos de minería de datos 1r, PRISM e ID3 a PostgreSQL*, JISTEM-Journal of Information Systems and Technology Management **10** (2013), no. 2, 389–406.

[2] Hanen Borchani, Gherardo Varando, Concha Bielza, and Pedro Larrañaga, *A survey on multi-output regression*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **5** (2015), no. 5, 216–233.

[3] Pedro Manuel Estrada Jiménez, Jorge Luis Ramírez de la Ribera, Danis Manuel Verdecia Acosta, and Yolanda Soler Pellicer, *Aplicación de la minería de datos en la estimación de componentes fotoquímicos (Original)*, Roca. Revista científico-educacional de la provincia Granma **15** (2019), no. 2, 177–186.

[4] Ron Kohavi et al., *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Ijcai, vol. 14, 1995, pp. 1137–1145.

[5] Sergio Valero Orea, Alejandro Salvador Vargas, and Marcela García Alonso, *Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos*, Ene **779** (2005), no. 73, 33.

[6] Oscar Reyes, Alberto Cano, Habib M Fardoun, and Sebastián Ventura, *A locally weighted learning method based on a data gravitation model for multi-target regression*, International Journal of Computational Intelligence Systems **11** (2018), no. 1, 282–295.

[7] Mariño Rivero and Adis Perla, *GMLKNN: modelo basado en instancias para el aprendizaje multi-etiqueta utilizando la distancia VDM*, Ph.D. thesis, Universidad Central Marta Abreu de Las Villas. Facultad de Matemática, 2015.

[8] Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Aikaterini Vrekou, and Ioannis Vlahavas, *Multi-target regression via random linear target combinations*, Joint european conference on machine learning and knowledge discovery in databases, Springer, 2014, pp. 225–240.

[9] Eleftherios Spyromitros-Xioufis, Grigorios Tsoumakas, William Groves, and Ioannis Vlahavas. Multi-label classification methods for multi-target regression. *arXiv preprint arXiv:1211.6581*, pages 1159–1168, 2012.

Sobre el/los autor/es:
Nombre: Pedro Manuel Estrada Jiménez
Correo electrónico: pestradaj@udg.co.cu
Institución: Departamento de Ciencias Básicas e Informática Aplicada, Universidad de Granma, Bayamo, Granma, Cuba.

Nombre: Pedro Jorge Noguera López
Correo electrónico: informatica@tramaote.co.cu
Institución: Departamento de Redes, Empresa Comercializadora de Tabaco en Rama "La Vega", UEB Camagüey Oriente, Bayamo, Granma, Cuba.

Nombre: Raúl Recio Avilés
Correo electrónico: rrecioa@udg.co.cu
Institución: Departamento de Ciencias Básicas e Informática Aplicada, Universidad de Granma, Bayamo, Granma, Cuba.