Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 184-193

Tipo de artículo: Artículo original
Temática: Reconocimiento de patrones
Recibido: 15/05/2016 | Aceptado:01/09/2016

Regresión lineal local con reducción de rango para problemas de predicción con salidas compuestas

Local linear regression with reduce rank for multitarget regression

Héctor R. González[1*], Carlos Morell[2], Antonio Blanco[1]

[1]Universidad de las Ciencias Informáticas. La Habana, Cuba. {adblanco}@uci.cu
[2]Universidad Central de las Villas "Marta Abreu" (UCLV), Villa Clara, Cuba. cmorellp@uclv.edu.cu

[*]Autor para correspondencia: hglez@uci.cu

Resumen

El propósito de este trabajo es estudiar el algoritmo de regresión con pesado local LWR para su adaptación a problemas de predicción con salidas compuestas. La idea de estimar los parámetros de la regresión lineal multivariada, mediante la descomposición en valores singulares, muestra una solución estable que permite manejar, a través del rango, la capacidad predictiva del algoritmo. Los experimentos desarrollados muestran que el algoritmo LWR es competitivo en el contexto del aprendizaje local basado en instancia. Los resultados obtenidos son un punto de partida para la adaptación de este enfoque, usando diversas estrategias que tomen en cuenta la interdependencia entre las variables de salidas.

Palabras clave: KNN, LWR, Regresión Lineal, Predicción con salidas compuestas, Aprendizaje con múltiples salidas

Abstract

The purposes of this work is to study Local Weighted Regression LWR algorithm for multi-target prediction problems. The idea to estimate, the parameters of multivariate linear regression trough to singular value decomposition and reduced rank show stable solution to drive the predicting performance of this algorithm. The experimental results show that LWR is a competitive algorithm, in context to instance based learning, for multi-target prediction problems. The preliminary results are the started point for futures adaptations, to this algorithms, take into account the interdependency between outputs variables.

Keywords: *KNN, LWR, Linear Regression, Multi-target Regression, Multi-Output Learning*

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 184-193

Introducción

Diversos problemas en reconocimiento de patrones y aprendizaje automático, han permitido la evolución de estas disciplinas a nuevos enfoques no convencionales de predicción o clasificación. Entre los métodos no convencionales que mayor desarrollo han tenido en los últimos años se encuentra la predicción estructurada (*Structured Prediction*) (Bakir et al., 2007; Pugelj & Džeroski, 2011), cuya diferencia con los métodos convencionales radica en que, la variable de salida, es modelada como una estructura de datos compleja. Estas estructuras, para la variable de salida, pueden ser modeladas como Grafos, Jerarquías, Secuencias, Cadenas de texto o Vectores, dependiendo del tipo de problema que se desea estudiar. En particular, aquellos problemas que se componen por vectores en la variable de salidas son conocidos como problemas de clasificación multietiqueta (Cuando son vectores con valores binarios) o predicción con salidas compuestas (Cuando las salidas están formadas por vectores reales)(Borchani, Varando, Bielza, & Larrañaga, 2015). Diversas aplicaciones han sido estudiadas en el ámbito de la predicción con salidas compuestas: en el campo de la quimiometría para evaluar diversos parámetros de calidad de agua (Džeroski, Demšar, & Grbović, 2000), en el modelado de los componentes de sistemas ecológicos (D Kocev, 2009) o para la venta de pasajes de cierta aerolínea (Spyromitros-Xioufis, Tsoumakas, Groves, & Vlahavas, 2012).

Un artículo de revisión publicado recientemente sobre los problemas de predicción con salidas compuestas establece dos grandes categorías (Borchani et al., 2015) para este tipo de tarea: El primer enfoque contiene los métodos que transforman el problemas de salidas múltiples en diferentes problemas de predicción de una salida; mientras que el otro enfoque se basa en adaptar métodos clásicos conocidos de aprendizaje automático para estimar de manera simultánea las múltiples salidas. En ambas categorías los métodos pueden tener en cuenta o no la interdependencia entre las variables de salidas.

Una ausencia notable entre los métodos que han sido adaptados, lo constituye el de los *K*-Vecinos más Cercanos (KNN) (Cover & Hart, 1967) a pesar de su simplicidad y posibilidades de escalabilidad. Este método ha demostrado ser competitivo en la modelación de problemas convencionales de clasificación y regresión para diversos dominios de aplicación. En la literatura consultada sobre predicción con salidas compuestas sólo se encontró reportado el trabajo (Pugelj & Džeroski, 2011), que propone el algoritmo KNN-SP para tres tipos de tareas de predicción estructurada. La propuesta de este algoritmo no toma en cuenta la interdependencia entre las variables de salidas en ninguna fase de su funcionamiento. Una evaluación estadística de este algoritmo y su comparación con los principales métodos reportados en el estado del arte arrojó que el KNN-SP alcanzaba los peores resultados en el Ranking de Friedman, presentando diferencias significativas con aquellos algoritmos que explotan la interdependencia entre las variables de

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 184-193

salida (González Diez, Santos, Campos, & Morell Pérez, 2016). A pesar de estos resultados, se considera que el buen comportamiento del KNN puede ser mejorado si se emplea un enfoque local de regresión con pesado, usando la regla de los vecinos más cercanos, conocido como LWR (*Local Weighted Regression*) por sus siglas en inglés (Frank, Hall, & Pfahringer, 2002; Atkeson, Moore, & Schaal, 1997).

El propósito del presente trabajo es adaptar la regla de los *K*-Vecinos más Cercanos al problema de predicción con salidas compuestas mediante un enfoque de regresión local pesado (LWR). En particular, para cada objeto a recuperar se aprende un modelo de regresión lineal multivariado para predecir de manera simultánea cada salida real de esta instancia. Debido a que el problema de regresión posee mucho menos instancias que atributos predictores, se utiliza para su solución la descomposición en valores singulares (SVD por sus siglas en ingles), lo cual permite reducir el rango y solo tomar en cuenta aquellas variables predictoras que son representativas. Nuestro estudio nos ayuda a evaluar el comportamiento de este algoritmo para diferentes valores de rango.

Materiales y métodos

En el presente trabajo usaremos las siguientes definiciones y notaciones para describir nuestro problema de predicción con salidas compuestas.

Sea $\vec{x} = [x_1, \dots, x_p] \in \Re^p$, $\vec{y} = [y_1, \dots, y_q] \in \Re^q$ dos vectores aleatorios en el espacio de entrada y salida respectivamente. Cada instancia de entrenamiento queda escrita como $(\vec{x}^i, \vec{y}^i) \in \Re^p \times \Re^q$, y el correspondiente problema de predicción con salidas compuestas consiste en estimar un único modelo $h: \Re^p \to \Re^q$, de modo que, la desviación esperada entre los valores reales de la variable de salida y el valor predicho, usando el modelo, se minimicen para todas las variables de entrada.

La regla de los *K*-Vecinos más Cercanos para problemas convencionales de predicción, utiliza la media de los vecinos sobre el conjunto de atributos de entrada para estimar la variable de salida. Su generalización a problemas de predicción con salidas compuesta consiste en determinar el valor medio para cada salida en el conjunto de los *K*-Vecinos más Cercanos. Una generalización a esta regla emplea una función de pesos que depende de la distancia del conjunto de vecinos a la instancia a recuperar. De manera formal, para una instancia $(\vec{x}^j, \vec{y}^j)$ con valores de la variable de salida desconocido y sus vecinos más cercanos $N_K(\vec{x}^j)$ para la variable de entrada $\vec{x}^j$ se estiman los valores de salida $\vec{y}^j$ según (1).

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 184-193

$$\hat{y}^j = \frac{\sum_{\vec{x}^i \in N_K(\vec{x}^j)} w(d(\vec{x}^i, \vec{x}^j)) y^i}{\sum_{\vec{x}^i \in N_K(\vec{x}^j)} w(d(\vec{x}^i, \vec{x}^j))} \quad (1)$$

Donde, $w(d(\vec{x}^i, \vec{x}^j))$ indica la función de pesos que pondera los vecinos recuperados a partir de los valores de distancias.

En el algoritmo de regresión local pesado (LWR) se estiman, para cada instancia a recuperar, un modelo de regresión lineal multivariada sobre el conjunto de los *K*-Vecinos más Cercanos y estimar cada vector de salida. Este modelo puede ser expresado según la función lineal:

$$\hat{y}^j = \hat{A}_{p \times q} \vec{x}^j + \hat{b}_{1 \times q} \quad (2)$$

Para estimar el modelo de regresión lineal multivariado se trata de minimizar el error cuadrático que se comete al estimar cada variable de salida del conjunto de entrenamiento con relación al valor real de dicha variable. En particular, LWR introduce una función de peso, que depende de la distancia entre el objeto a recuperar y sus vecinos, para modelar un problema de optimización convexo en la forma.

$$L(A, b) = \sum_q \sum_{\vec{x}^i \in N_K(\vec{x}^j)} \left({y_q}^i - A_{:,q} \vec{x}^i - b_q \right)^2 w(d(\vec{x}^i, \vec{x}^j)) \quad (3)$$

Donde $A_{:,q}$ indica la columna q de dicha matriz. Si realizamos la transformación $Z = \sqrt{w(d(\vec{x}^i, \vec{x}^j))} \vec{x}^i$ y $G = \sqrt{w(d(\vec{x}^i, \vec{x}^j))} \vec{y}^i$, obtenemos un problema de optimización similar a la regresión lineal multivariada sin pesos en las nuevas variables y cuya solución para estimar la matriz de regresión del modelo lineal quedará expresada en su forma normal como:

$$\hat{A}_{p+1 \times q} = (Z^T Z)^{-1} Z^T G \quad (4)$$

En el problema de LWR para la generalidad de las bases de datos se cumple que, la cardinalidad del conjunto de vecinos es mucho menor que el número de atributos $\left\| N_K(\vec{x}^j) \right\| \ll p$, de ahí que $Z^T Z$ es una matriz singular en la mayoría de los casos y por lo tanto no inversible. Este problema puede ser resuelto introduciendo un factor de contracción como regularizador, sin embargo, algunos autores demuestran que la solución a este problema, utilizando el cálculo de la inversa, puede ser inestable. Una alternativa para dar solución al problema de no invertir esta matriz, puede desarrollarse mediante la descomposición de la matriz *Z* en valores singulares (SVD) (Golub & Reinsch, 1970; Kleibergen & Paap, 2006).

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 184-193

En general una matriz arbitraria $A_{p\times q}$, admite una descomposición en la forma $A = U_{p\times p}SV^T{}_{q\times q}$ Las matrices *U* y *V* son ortogonales y conforman el conjunto de vectores singulares de *A*, mientras que *S* es una matriz diagonal de rango $r \leq \min(p, q)$, con valores positivos ordenados de mayor a menor.

Para estimar los valores de la matriz *A*, correspondiente al modelo de regresión lineal multivariado, se puede demostrar, aprovechando las propiedades de las matrices ortogonales de la descomposición, que la estimación de esta se determina por la expresión (5):

$$\hat{A} = US^{-1}V^TG \qquad (5)$$

Donde, $S^{-1} = diag(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_r}, 0, \dots, 0)$. Para valores pequeños de σ_i su inversa puede crecer infinitamente por lo que estos valores pueden truncarse a cero y por consiguiente reducir el rango.

Resultados y discusión

En esta sección, describiremos la configuración de los experimentos diseñados y discutiremos los principales resultados relacionados con el algoritmo propuesto.

Como un primer elemento, se presentan los detalles técnicos relacionados con las bases de datos, parámetros de ajustes de los algoritmos y la implementación. Posteriormente, se presenta una comparación entre el enfoque local de regresión pesado y el clásico KNN-SP sobre ocho bases de datos públicas disponible para problemas de predicción con salidas compuestas.

En la experimentación se emplean diferentes cantidades del número de vecinos en la predicción final k_p, cuyos valores se reportan en los resultados experimentales para los algoritmos evaluados. Para los experimentos reportados en este artículo, los valores del número de vecinos se encuentran en el intervalo [7,37] donde solo se evaluaron los valores impares de este intervalo.

Para el algoritmo KNN-SP, se emplea en la predicción el pesado por el inverso de la distancia y por la función Gaussiana de la distancia como se muestra en la expresión (6), mientras que las mismas funciones de pesos son empleadas en el algoritmo LWR. Para el LWR se reportan los resultados para varios valores de rango y su influencia en la predicción.

$$w_g(\vec{u}, \vec{v}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{d(\vec{u},\vec{v})}{\sigma}}, \qquad w_i(\vec{u}, \vec{v}) = \frac{1}{1+d(\vec{u},\vec{v})} \qquad (6)$$

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 184-193

En la experimentación, se emplea la validación cruzada con 10-folds para cada base de datos, excepto las bases de datos de mayor dimensión (más de 1000 instancias), que han sido divididas en entrenamiento y prueba. La Tabla (1) resume las principales características de estas bases de datos (Borchani et al., 2015; Spyromitros-Xioufis, Tsoumakas, William, & Vlahavas, 2014). El procedimiento de validación cruzada ha sido integrado en el paquete de software MULAN (Tsoumakas, Spyromitros-Xioufis, Vilcek, & Vlahavas, 2011).

Tabla 1 Bases de datos utilizada en la experimentación y la información asociada a cada una de ellas.

Dataset	Instancias	Atributos de entrada	Atributos de salida
waterquality	1060	16	14
jura	359	15	3
enb	768	8	2
slump	103	7	3
osales	639	413	12
scpf	1137	23	3
edm	154	16	2
atp1d	201/136	411	6

Como en trabajos similares, nosotros empleamos para medir la efectividad de los algoritmos, la métrica del error cuadrático medio relativo (RRMSE), determinando su valor medio sobre el conjunto de salidas de la siguiente manera:

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

Para comparar estadísticamente los resultados, en el presente trabajo empleamos la prueba de Wilcoxon con rangos de signos para comparar el par de algoritmos evaluados en la experimentación, siguiendo las recomendaciones realizadas en (Demšar, 2006) y sus extensiones propuestas en (García, Fernández, Luengo, & Herrera, 2009).

Estas bases de datos se relacionan en la tabla 1 donde la primera columna indica el nombre de la base de datos, la segunda se refiere al número de observaciones de la base de datos. En esta columna además se indica si los datos están particionados en datos de prueba (test) y datos de entrenamiento (train) o si la base de datos contiene todos los datos los cuales deben ser particionados para los experimentos. Para particionar los datos se utiliza un método de validación cruzada el cual se conoce como *K-Fold Cross Validation* (lo señalamos como CV). La 3ra y 4ta columna indica la cantidad de atributo en el espacio de entrada y salida respectivamente. La última columna brinda una breve explicación del dominio de aplicación en el cual fueron colectados los datos.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 184-193

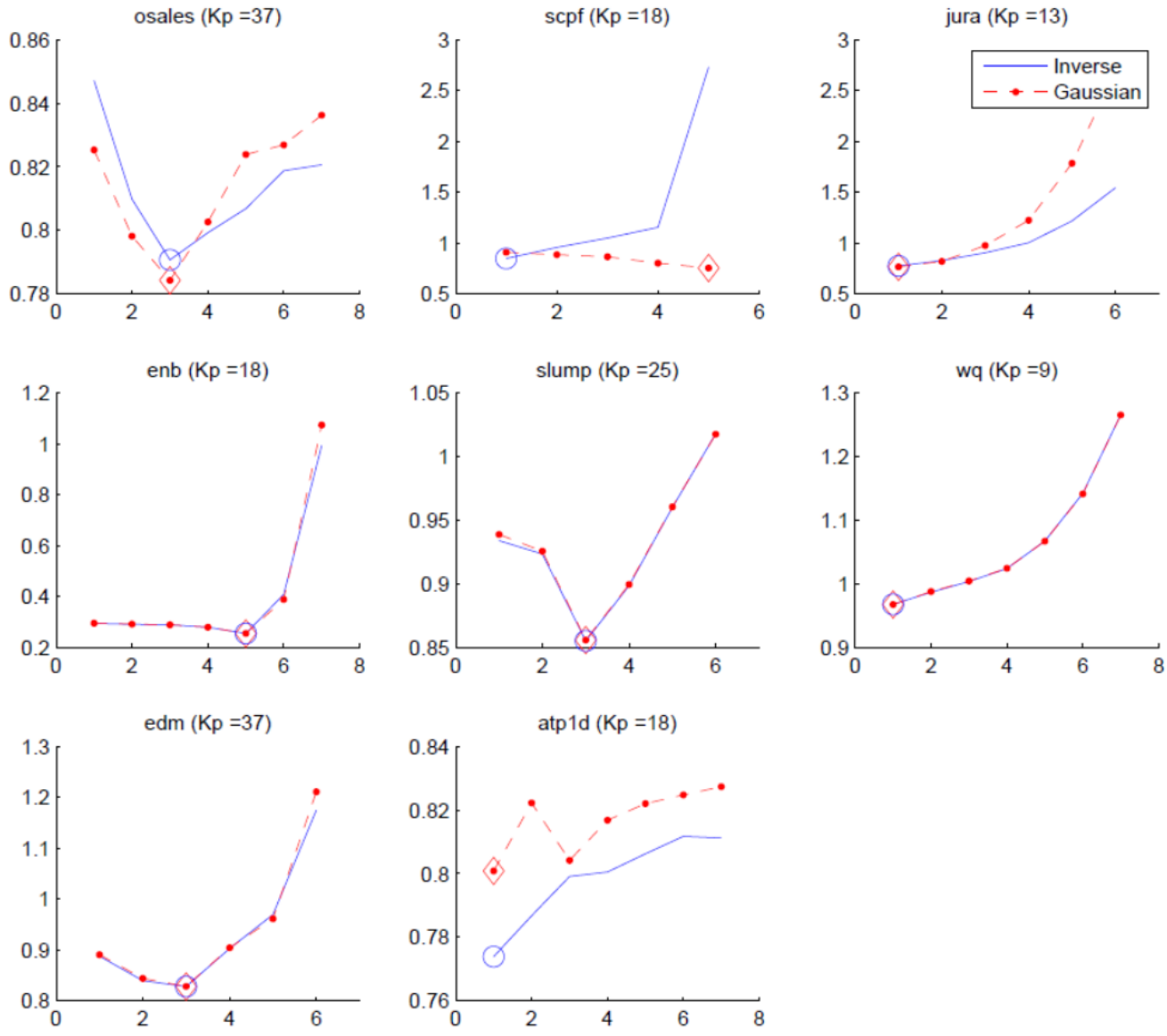

Figura 1. Valores del aRRMSE correspondiente a diferentes valores de rango r, usando pesado Inverso y Gaussiano. El mejor valor de rango es marcado en cada gráfica.

La Figura 1 contiene los resultados del aRRMSE versus el rango de la descomposición en valores singulares *r*, para las 8 bases de datos consideradas en el trabajo. El mejor valor del número de vecinos para la predicción k_p es reportado en la gráfica. Adicionalmente, los mejores resultados son señalizados en las gráficas con un círculo y un diamante de colores azul y rojo respectivamente según el tipo de función de peso. Como se puede apreciar con un

valor del rango relativamente pequeño se obtienen los mejores resultados. Los valores de rango superiores a 7 no fueron reportados en la gráfica pues estos arrojaban resultados muy desequilibrados y distorsionaba la representación visual de la misma.

En la tabla 2, se detallan los mejores resultados comparándolos con los alcanzados para el algoritmo KNN-SP. La aplicación de la prueba de Wilcoxon entre las variantes de LWR estudiadas y el KNN-SP, arrojó que el valor crítico ($p = 0.3828$) es mayor que 0.05 para ambos casos por lo que se acepta la Hipótesis nula de esta prueba y por consiguiente afirmamos que no existen diferencias significativas entre ambos algoritmos con un 95 % de confianza.

Tabla 2 aRRMSE obtenidos para las variantes de LWR y KNN-SP en las 8 bases de datos estudiadas.

Dataset	LWR_{Inv}	LWR_{Gau}	KNN-SP
waterquality	0.967	0.967	0.948
jura	0.774	0.765	0.725
enb	0.254	0.254	0.283
slump	0.855	0.856	0.760
osales	0.791	0.784	1.012
scpf	0.846	0.752	0.958
edm	0.827	0.828	0.836
atp1d	0.774	0.801	0.948

Conclusiones

Los resultados alcanzados en este trabajo, nos indican que los modelos basados en regresión local son competitivos en el contexto de la predicción con salidas compuestas. Por otra parte, los valores de regresión se ven fuertemente afectados por el cálculo de la inversa en el modelo de regresión. En caso de emplear descomposición en valores singulares los resultados de predicción son fuertemente dependientes del manejo del rango.

El estudio y evaluación de los algoritmos de regresión local con pesado LWR para problemas de predicción con salidas compuestas, es un punto de partida a futuras investigaciones. Estas direcciones de trabajo se enmarcan en adaptar estos enfoques al tratamiento de la interdependencia entre las variables de salidas empleando diversas estrategias.

Referencias

ATKESON, C. G., MOORE, A. W., & SCHAAL, S. (1997). Locally weighted learning for control. In *Lazy learning* (pp. 75–113). Springer.

BAKIR, G. H., HOFMANN, T., SCHÖLKOPF, B., SMOLA, A. J., TASKAR, B., & VISHWANATHAN, S. V. N. (2007). *Predicting Structured Data (Neural Information Processing)*. The MIT Press.

BORCHANI, H., VARANDO, G., BIELZA, C., & LARRAÑAGA, P. (2015). A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *5*(5), 216–233.

COVER, T., & HART, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, *13*(1), 21–27.

DEMŠAR, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, *7*, 1–30.

D KOCEV, P. G., S DZEROSKI, M D WHITE, G R NEWELL. (2009). Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. *Ecological Modelling 220*.

DŽEROSKI, S., DEMŠAR, D., & GRBOVIĆ, J. (2000). Predicting chemical parameters of river water quality from bioindicator data. *Applied Intelligence*, *13*(1), 7–17.

FRANK, E., HALL, M., & PFAHRINGER, B. (2002). Locally weighted naive bayes. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence* (pp. 249–256). Morgan Kaufmann Publishers Inc.

GARCÍA, S., FERNÁNDEZ, A., LUENGO, J., & HERRERA, F. (2009). A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing*, *13*(10), 959–977.

GOLUB, G. H., & REINSCH, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, *14*(5), 403–420.

GONZÁLEZ DIEZ, H. R., SANTOS, G., CAMPOS, F., & MORELL PÉREZ, C. (2016, July 13). Evaluación del algoritmo KNN-SP para problemas de predicción con salidas compuestas. *Revista Cubana de Ciencias Informáticas*, *10*(3), 119–129.

KLEIBERGEN, F., & PAAP, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, *133*(1), 97–126.

PUGELJ, M., & DŽEROSKI, S. (2011). Predicting structured outputs k-nearest neighbours method. In *Discovery Science* (pp. 262–276). Springer.

SPYROMITROS-XIOUFIS, E., TSOUMAKAS, G., GROVES, W., & VLAHAVAS, I. (2012). multi-label classification methods for multi-Target Regression. *arXiv Preprint arXiv:1211.6581*.

SPYROMITROS-XIOUFIS, E., TSOUMAKAS, G., WILLIAM, G., & VLAHAVAS, I. (2014). Drawing Parallels between Multi-Label Classification and Multi-Target Regression. *arXiv Preprint arXiv:1211.6581v2*.

TSOUMAKAS, G., SPYROMITROS-XIOUFIS, E., VILCEK, J., & VLAHAVAS, I. (2011). Mulan: A java library for multi-label learning. *The Journal of Machine Learning Research*, *12*, 2411–2414.