Congreso Internacional en Ingeniería Electrónica. Mem. Electro 2015, Vol. 37, pp 292-297, Chihuahua, Chih. México
http://depi.itchihuahua.edu.mx/display/memorias_electro/MemoriaElectro2015
ISSN 1405-2172

REDUCCIÓN DE LA DIMENSIONALIDAD EN PROBLEMAS DE REGRESIÓN LINEAL

Blanco Vega Ricardo[1], De la Cruz Nava José Alfredo[2], Torres Knight Ricardo Ramón[3], Blanco Vega Humberto[4]
Universidad Autónoma de Chihuahua
Facultad de Ingeniería[123]
Circuito Universitario Campus II
Tel. (614) 4425024
Fax. (614) 4425000
Facultad de Educación Física y Ciencias del Deporte[4]
Cd. Universitaria, Apdo. postal 2-1585
C.P. 31009.
Tel. (614) 413-0433
Chihuahua, Chih., México
rblanco@uach.mx[1], alfredodelac@hotmail.com[2], rtorres@uach.mx[3], hblanco@uach.mx[4]

RESUMEN

El objetivo del presente trabajo es automatizar el aprendizaje de un modelo de regresión trascendente univariable y mejorar el modelo lineal obtenido con todos sus atributos originales, usando el mejor atributo para predecir la clase en un conjunto de datos. Luego realizando tranformaciones sobre el mejor atributo y la clase se aprenden diferentes tipos de modelos de regresión: logarítmico, exponencial, potencia, lineal, cuadrático, y cúbico. Finalmente se selecciona el modelo que tiene el coeficiente de correlación más alto. Para incrementar la utilidad del modelo aprendido se construye automáticamente un programa computacional de predicción. Este programa ayuda a que cualquier persona que lo use actúe como experto estimando valores futuros. Se experimenta con 36 casos y se encuentra que en promedio se incrementa un 39% el coeficiente de determinación y la dimensionalidad se reduce en un 84%.

Palabras Clave: modelo de regresión, modelo lineal, modelo trascendente, modelo logarítmico, modelo exponencial, modelo potencia, modelo lineal, modelo cuadrático, modelo cúbico, predecir clase, coeficiente de correlación, sistema de predicción, reducción de la dimensionalidad.

ABSTRACT

The aim of this paper is to automate the learning model of univariate regression transcendent better than the linear model obtained with all their original attributes. Using the best attribute to predict the class in a dataset. Then performing transformations on the best attribute and the class different types of regression models are learned: logarithmic, exponential, power, linear, quadratic, and cubic. Finally the model that has the highest correlation coefficient is selected. To increase the usefulness of the model learned a computational prediction program is built automatically. This program helps anyone who can act as an expert use estimating future values. It is experimented with 36 cases and found that on average 39% increases the coefficient of determination and the dimensionality is reduced by 84%.

Keywords: regression model, linear model, transcendent model, logarithmic model, exponential model, power model, linear model, quadratic model, cubic model to predict class correlation coefficient prediction system, dimensionality reduction.

1. INTRODUCCIÓN

Los sistemas de predicción usando regresión lineal son de gran ayuda para la toma de decisiones en cualquier organización o empresa. El proceso de obtener un modelo de regresión de calidad es tardado y requiere de conocimiento experto. La calidad del modelo se ve disminuida por el uso de muchos atributos o datos de entrada. Para incrementar la calidad del modelo a obtener es necesario reducir la dimensión del problema eliminando atributos sin perder coeficiente de determinación [7]. De esta manera el modelo permitirá realizar predicciones más rápidas, más comprensión de la operación de la predicción y con un costo menor en la adquisición de datos.

El objetivo de la reducción de la dimensionalidad en la regresión es encontrar en una amplia gama de datos un regresor eficiente reduciendo la cantidad de entradas. De esta manera no solo el

modelo queda más comprensible y práctico sino que es posible que sea más precisa su predicción.

Hay diferentes técnicas para reducir la dimensionalidad, por ejemplo, en la Universidad de Rutgers se utilizó la regresión con un operador de covarianza inverso. La técnica es eficiente y descubren espacios subcentrales fiables y robustos para eliminar la alta dimensionalidad en los datos [5].

Para reducir la dimensionalidad en los problemas de regresión se usa el análisis de componente independiente [6]. Se eliminan variables irrelevantes o redundantes que complican el proceso de aprendizaje. Se utiliza el error mínimo cuadrático como parámetro de calidad del modelo aprendido.

La reducción de la dimensionalidad se realiza considerando dimensiones en el conjunto de datos. Las dimensiones se almacenan en "ramas", y se realizar manipulaciones para ver su importancia [11].

En los estudios de eficacia clínica el resultado es a menudo influenciado por múltiples factores causales, como las drogas, el incumplimiento, la frecuencia de la consulta, y muchos más factores. Los médicos buscan la reducción de estos múltiples factores para estima con buena probabilidad la eficacia clínica [3].

En [4] se propone un método de selección de atributos tomando como base la regresión lineal y se obtienen modelos de mayor calidad.

Como hemos visto las técnicas de reducción de dimensionalidad son diversas y diferentes, nuestra estrategia propone utilizar la técnica de regresión lineal simple y usando transformaciones en el atributo más importante y la clase poder encontrar un modelo con alta calidad que supere al obtenido con todos los atributos originales.

2. DESARROLLO

Uno de los problemas principales en los sistemas expertos o los sistemas de predicción es el cuello de botella en la adquisición del conocimiento, debido principalmente a que muchos expertos no expresan su conocimiento en forma de reglas claras e inequívocas. Los expertos emplean generalmente reglas explícitas y reglas empíricas (sin ningún fundamento lógico). Incluso en caso de que el experto describa todo su conocimiento, esto representa un alto esfuerzo, es un proceso donde se desperdicia mucho tiempo y es difícil de mantener. Además, el resultado es a veces un modelo trascrito que no se aplica de una forma completamente automatizada ya que todavía presenta una cierta ambigüedad.

En algunos ámbitos (por ejemplo, el comercial) el problema se plantea como la estimación futura de un valor numérico en función de sus hechos históricos. Los hechos están descritos por una serie fija de atributos y por un valor dependiente numérico. Los modelos creados por los expertos predicen el valor dependiente de acuerdo con el resto de los atributos. Esta estructura de modelo es similar a la de los modelos de regresión en aprendizaje computacional, donde tenemos casos expresados a partir de varias variables de entrada y una variable de salida.

En este caso, el problema de la adquisición del conocimiento se resuelve por técnicas de aprendizaje computacional usando Minería de Datos. La idea es que preguntemos a un experto para obtener los datos a partir de los cuales construir nuestro modelo. O simplemente tenemos un conjunto de datos que almacenan el comportamiento histórico de una empresa.

En cualquiera de los dos casos anteriores partiremos de un conjunto de datos expresado en un archivo con formato csv (archivo de texto separado por comas). Estos datos servirán de materia prima o vista minable. La tabla 1 muestra una lista de datos experimentales usados en éste trabajo.

2.1 VALIDACIÓN EXPERIMENTAL

Para llevar a cabo este sistema computacional se automatiza la metodología propuesta por Blanco et. al. [2]. Se procede primero encontrando el mejor algoritmo de regresión y luego usando ese algoritmo encontrar el mejor modelo, para luego usarlo en la creación de un sistema de predicción.

Nuestro trabajo se centra en considerar a la regresión lineal como base para encontrar el mejor algoritmo. Y nuestra tarea es reducir al máximo los atributos independientes. Para ello se selecciona el mejor atributo. El mejor atributo es el que posee el coeficiente de correlación más alto.

Tabla 1. Conjuntos de datos utilizados en los experimentos. Donde los nombres de las columnas se refieren a propiedades del conjunto de datos a saber: No. es un número de identificación, datos es su nombre, Atr. es el número de atributos totales sin considerar la clase, Num. es la cantidad de atributos numéricos, Nom. es la cantidad de atributos nominales, Tam. es el número total de registros y Fal. es si existen valores perdidos o faltantes.

No.	Datos	Atr.	Num.	Nom.	Tam.	Fal.
1	auto93	22	16	6	93	Sí
2	autoHorse	25	17	8	205	Sí
3	autoMpg	7	4	3	398	Sí
4	autoPrice	15	15	0	159	No
5	baskball	4	4	0	96	No
6	bodyfat	14	14	0	252	No
7	bolts	7	7	0	40	No
8	breastTumor	9	1	8	286	Sí
9	cholesterol	13	6	7	303	Sí
10	cleveland	13	6	7	303	Sí
11	cloud	6	4	2	108	No
12	cpu	6	6	0	209	No
13	detroit	13	13	0	13	No
14	echoMonths	9	6	3	130	Sí
15	elusage	2	1	1	55	No
16	fishcatch	7	5	2	158	Sí
17	fruitfly	4	2	2	125	No
18	gascons	4	4	0	27	No
19	housing	13	12	1	506	No
20	hungarian	13	6	7	294	Sí
21	longley	6	6	0	16	No
22	lowbwt	9	2	7	189	No
23	mbagrade	2	1	1	61	No
24	meta	21	19	2	528	Sí
25	pbc	18	10	8	418	Sí
26	pharynx	11	1	10	195	Sí
27	pollution	15	15	0	60	No
28	pwLinear	10	10	0	200	No
29	quake	3	3	0	2178	No
30	schlvote	5	4	1	38	No
31	sensory	11	0	11	576	No
32	servo	4	0	4	167	No
33	sleep	7	7	0	62	Sí
34	strike	6	5	1	625	No
35	veteran	7	3	4	137	No
36	vineyard	3	3	0	52	No

Se utiliza la estrategia de “a los datos hay que torturarlos hasta que confiesen” [8], es decir, el mejor atributo y la clase se tranforman para crear diferentes tipos de modelos. Las transformaciones se resumen en la tabla 2.

Tabla 2. Transformaciones realizadas a los datos para obtener los diferentes modelos y solamente utilizar regresión lineal simple.

Modelo	Transformación a y (clase)	Transformación a x (mejor atributo)
Lineal	ninguna	ninguna
Logarítmica	ninguna	ln(x)
Exponencial	ln(y)	ninguna
Potencial	ln(y)	ln(x)
Cuadrático	ninguna	adicionar x^2
Cúbico	ninguna	adicionar x^2 y x^3

A continuación se describen las diferentes transformaciones utilizadas.

Modelo Lineal

Calcula el número mínimo de cuadrados en una línea utilizando la siguiente ecuación: y=mx+b, donde m es la pendiente y b es la intersección. Una línea de tendencia lineal es una línea recta que se ajusta perfectamente y que se utiliza con conjuntos de datos lineales simples. Los datos son lineales si la trama de los puntos de datos se parece a una línea. Una línea de tendencia lineal normalmente muestra que algo aumenta o disminuye a un ritmo constante.

Modelo Logarítmico

Calcula el número mínimo de cuadrados mediante puntos utilizando la siguiente ecuación: y=c ln (x)+b, donde c y b son constantes y ln es el logaritmo natural (neperiano). Una línea de tendencia logarítmica es una línea curva que se ajusta perfectamente y que se utiliza cuando el índice de cambios de los datos aumenta o disminuye rápidamente y después se estabiliza.

Modelo Exponencial

Calcula el número mínimo de cuadrados mediante puntos utilizando la siguiente ecuación: $y=c\ e^{bx}$, bx son exponentes de e, donde c y b son constantes y e es la base del logaritmo neperiano. Una línea de tendencia exponencial es una línea curva que se utiliza cuando los valores de los datos aumentan o disminuyen a intervalos cada vez mayores. No es posible crear una línea de tendencia exponencial si los datos contienen valores cero o negativos.

Modelo Potencial

Calcula el número mínimo de cuadrados mediante puntos utilizando la siguiente ecuación: $y=c\ x^b$, b es exponente de x, donde c y b son constantes. Una línea de tendencia de potencia es una línea curva utilizada con conjuntos de datos que comparan medidas que aumentan a un ritmo concreto; por ejemplo, la aceleración de un automóvil de carreras a intervalos de un segundo. No es posible crear una línea de tendencia de potencia si los datos contienen valores cero o negativos.

Modelo Cuadrático

Calcula el número mínimo de cuadrados mediante puntos utilizando la siguiente ecuación: $y=c_0 + c_1x + c_2x^2$, donde c_i son constantes.

Modelo Cúbico

Calcula el número mínimo de cuadrados mediante puntos utilizando la siguiente ecuación: $y=c_0 + c_1x + c_2x^2 + c_3x^3$, donde c_i son constantes.

2.2 HERRAMIENTA CASE

En la automatización se utiliza Java y WEKA [10]. WEKA es un conjunto de paquetes para realizar Minería de Datos programada en Java creada en la Universidad de Waikato, Nueva Zelanda.

El programa realiza las siguientes tareas, ver las opciones de menú en la figura 1:

1. Abrir Datos. Carga los datos a memoria desde un archivo arff o csv.
2. Obtener el Atributo con Mayor R.
3. Determinar el Mejor Modelo. Comparando los modelos: lineal, logarítmico, exponencial, potencia, cuadrático y cúbico.
4. Crear el programa de Regresión. Se utiliza el mejor modelo obtenido.

Opcionalmente se tiene la posibilidad de ejecutar de forma conjunta todas las opciones. La última opción es para salir del programa.

El mejor atributo se obtiene al hacer competir a todos los atributos con una validación cruzada de 10 divisiones usando la configuración predeterminada de algoritmo de regresión lineal y comparando el parámetro de calidad R, es decir, el mejor atributo es aquel que genera el modelo con el mayor R.

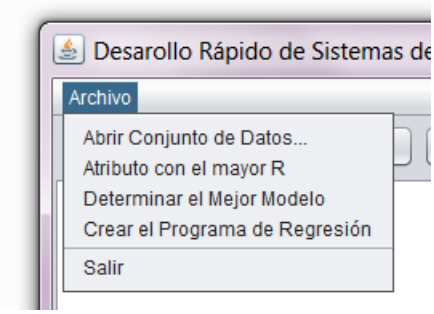

Figura 1. Opciones del CASE RADRU.

La selección de los atributos se realiza con la ordenación de los R generados al utilizar solamente el atributo y la clase. Es decir, se crea un conjunto de datos para cada atributo y la clase, luego se crea el modelo y finalmente se realiza una evaluación cruzada de 10 para obtener el R.
Para obtener el mejor modelo se utiliza el mejor atributo y la clase, se crea un modelo y se evalúa para obtener su R^2, después se realizan las transformaciones descritas en la tabla 2 para obtener los otros modelos. Se selecciona el mejor R^2 para definir el mejor modelo.

Ese mejor modelo se utiliza para crear el sistema de regresión. El sistema de regresión está constituido por los archivos: el programa (programa.jar), la librería de WEKA (weka.jar), el mejor modelo (archivo .model) y el conjunto de datos con una sola instancia o registro (archivo .arff). La herramienta CASE crea una carpeta llamada programa y en ella copia los 4 archivos mencionado anteriormente. El usuario debe moverse a la carpeta programa y ejecutar el archivo programa.jar. El pre-requisito para poder correr el CASE y el programa.jar es tener instalada la máquina virtual de Java.

3. RESULTADOS

La interfaz gráfica de usuario de la herramienta CASE se observa en la figura 2.

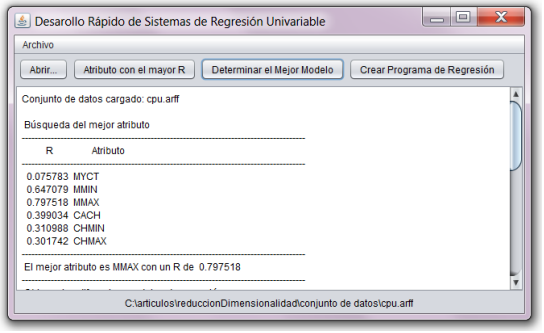

Figura 2. Interfaz gráfica de usuario del CASE Desarrollo Rápido de Sistemas de Regresión Univariable.

Para utilizar el programa primero se carga un conjunto de datos y luego se ejecuta en cada una de las tareas o todas en conjunto. A continuación se muestra el resultado de una ejecución del CASE usando el conjunto de datos cpu.

Conjunto de datos cargado: cpu.arff

Búsqueda del mejor atributo

R Atributo

0.075783 MYCT
0.647079 MMIN
0.797518 MMAX
0.399034 CACH
0.310988 CHMIN
0.301742 CHMAX

El mejor atributo es MMAX con un R de 0.797518

Obtener los diferentes modelos de regresión.

No. Tipo R

---

1 Todos 0.8612
2 Lineal 0.7975
3 Exponencial 0.8436
4 Logarítmico 0.3670
5 Potencia 0.7573
6 Cuadrática 0.8934
7 Cúbica 0.8892

El mejor modelo con R=0.8934064832373946 Cuadrática

Linear Regression Model

class =

0.0023 * MMAX +
0 * x2 +
17.6821

Creando el directorio programa.
Creando el directorio de la librería.
Copiando el conjunto de datos.
Guardando el modelo.
Copiando el programa jar.
Copiando la librería WEKA.
El programa de predicción se ha creado en la carpeta programa.
Para ejecutarlo doble click sobre el archivo programa.jar

El sistema computacional se evalúa con pruebas en 36 conjuntos de datos del repositorio UCI [1] los resultados se muestran en la tabla 3.

En la tabla 3 los nombres de las columnas se refieren a propiedades del modelo obtenido a saber: No. es un número de identificación del conjunto de datos de acuerdo a la tabla 1, Mejor Atr. Es el mejor atributo con mayor R^2, las siguientes columnas son los coeficientes de correlación de los modelos de regresión obtenido con todos los atributos, lineal, logarítmico, exponencial, potencia, cuadrático, cúbica y el mejor de los modelos univariados. La columna de mejor modelo presenta el nombre del tipo de modelo que fue el mejor en R^2 y finalmente la columna del porciento de ganancia del mejor modelo contra el modelo con todos los atributos en R^2.

Tabla 3. Resultados de las pruebas.

No.	Mejor Atr.	R2							Mejor Modelo	% Ganancia
		Todos	Lineal	Logarítmica	Exponencial	Potencia	Cuadrática	Cúbica	Mejor		
1	Horsepower	0.6653	0.5768	0.6446	0.5721	0.7058	0.5688	0.5727	0.7058	Potencia	6.1
2	engine-size	0.8913	0.6464	0.6093	0.6539	0.6564	0.6589	0.6309	0.6589	Cuadrática	-26.1
3	weight	0.8547	0.6884	0.7644	0.7097	0.7626	0.7115	0.7113	0.7644	Exponencial	-10.6
4	curb-weight	0.7626	0.7901	0.8364	0.7483	0.8338	0.8109	0.8080	0.8364	Exponencial	9.7
5	time_played	0.3722	0.2974	0.3002	0.2830	0.2925	0.2892	0.2824	0.3002	Exponencial	-19.3
6	Density	0.9726	0.9749	0.8464	0.9761	0.8413	0.9771	0.9777	0.9777	Cúbica	0.5
7	TIME	0.8439	0.7225	0.5861	0.7478	0.7139	0.7528	0.7603	0.7603	Cúbica	-9.9
8	inv-nodes	0.0949	0.0442	0.0227	0.0278	0.0367	0.0511	0.0462	0.0511	Cuadrática	-46.2
9	age	0.0252	0.0289	0.0269	0.0313	0.0299	0.0258	0.0259	0.0313	Logarítmica	24.2
10	ca	0.5034	0.2576	0.0867	0.0348	0.0433	0.2442	0.2333	0.2576	Lineal	-48.8
11	SC	0.8555	0.7158	0.5612	0.4689	0.7290	0.6808	0.6540	0.7290	Potencia	-14.8
12	MMAX	0.8612	0.7975	0.8436	0.3670	0.7573	0.8934	0.8892	0.8934	Cuadrática	3.7
13	FTP	0.4476	0.6728	0.6440	0.6411	0.6206	0.6730	0.6850	0.6850	Cúbica	53.0
14	still_alive	0.5134	0.4797	0.6032	0.1671	0.1088	0.4797	0.4797	0.6032	Exponencial	17.5
15	average_temperature	0.7510	0.7691	0.7984	0.8225	0.8047	0.8369	0.8369	0.8369	Cuadrática	11.4
16	LengthG	0.9237	0.8507	0.8171	0.7188	0.9437	0.8744	0.8811	0.9437	Potencia	2.2
17	SLEEP	0.0847	0.0847	0.0029	0.0575	0.0097	0.0837	0.0295	0.0847	Lineal	0.0
18	year	0.3114	0.8202	0.8058	0.8213	0.8071	0.9063	0.9062	0.9063	Cuadrática	191.0
19	LSTAT	0.7142	0.5348	0.6409	0.6585	0.6732	0.6331	0.6495	0.6732	Potencia	-5.7
20	exang	0.5145	0.3300	0.0000	0.0460	0.0000	0.3300	0.3300	0.3300	Lineal	-35.9
21	GNP	0.9548	0.9613	0.9592	0.9517	0.9559	0.9573	0.9505	0.9613	Lineal	0.7
22	LOW	0.5948	0.6046	0.5976	0.0776	0.0681	0.6046	0.6046	0.6046	Lineal	1.6
24	DS_Name	0.0617	0.0642	0.2675	0.0032	0.0082	0.0642	0.0642	0.2675	Exponencial	366.0
25	Z10	0.3553	0.1792	0.2024	0.1766	0.2088	0.1792	0.1672	0.2088	Potencia	-41.2
27	NDNW	0.5510	0.3945	0.3843	0.2352	0.2311	0.3945	0.3946	0.3945	Lineal	-28.4
28	a1	0.7432	0.4519	0.1780	0.0409	0.0882	0.4519	0.4519	0.4519	Lineal	-39.2
29	latitude	0.0028	0.0037	0.0037	0.0036	0.0037	0.0024	0.0065	0.0065	Cúbica	132.1
30	budget	0.0442	0.0412	0.1250	0.1548	0.2463	0.0028	0.0874	0.2463	Potencia	457.2
31	Flows	0.1495	0.0762	0.0782	0.0758	0.0758	0.0762	0.0762	0.0782	Exponencial	-47.7
32	pgain	0.7100	0.5712	0.6211	0.3173	0.3685	0.5712	0.5712	0.6211	Exponencial	-12.5
33	sleep_exposure_index	0.3319	0.3639	0.4541	0.3211	0.3751	0.3801	0.3439	0.4541	Exponencial	36.8
34	country	0.1675	0.1669	0.6115	0.0126	0.0105	0.1669	0.1683	0.6115	Exponencial	265.1
35	karnofsky	0.1390	0.1105	0.2964	0.0970	0.2635	0.0984	0.0863	0.2964	Exponencial	113.2
36	lugs_1989	0.5379	0.5204	0.3368	0.3568	0.1686	0.6192	0.6695	0.6695	Cúbica	24.5
	MediaModa	0.5090	0.4586	0.4581	0.3641	0.3952	0.4721	0.4712	0.5271	Exponencial	39.1

Para obtener mejores estadísticas se eliminaron los conjuntos de datos 23 con el peor resultado (una ganancia de -51.8%) y 26 con el mejor resultado (una ganancia de 4,191.9%).

En la tabla 4 se observan las frecuencias de los algoritmos ganadores ordenadas de mayor a menor frecuencia.

Tabla 4. Frecuencias de modelos ganadores.

Tipo	# Gano Modelos	# Gana a Todos	% Ganancia R^2
Exponencial	10	6	71.8
Cúbica	5	4	40.1
Cuadrática	5	3	26.8
Potencia	6	3	67.3
Lineal	7	2	-21.4
Logarítmica	1	1	24.2
Total	34	19	

El modelo que más veces gana es el exponencial. Existe una reducción de atributos en un promedio de 84% y una ganancia del 39.1% en R^2.

4. CONCLUSIONES

Se cumple con el objetivo de reducir los atributos de un problema de regresión lineal a un 84%. De 34 casos estudiados se observa que en 17 casos se gana. Respecto al coeficiente de determinación se observa en promedio que aumenta en un 39.1% en los casos que gana el modelo univariado. También se proporciona un programa que automatiza la obtención del modelo univariado. Estos resultados permiten a las empresas reducir el costo de la obtención de sus datos dado en un 56% de los casos es necesario solamente una variable para poder predecir.

El modelo que gana más veces y con un porcentaje mayor de R^2 es el exponencial.

La herramienta solamente tiene en consideración para la evaluación de la calidad el coeficiente de correlación o el coeficiente de determinación. Para la obtención de un sistema de regresión de alta calidad es esencial el trabajo del experto para validar las características subjetivas del conocimiento, como son la utilidad y la novedad.

Como trabajo futuro se pretende buscar el mejor algoritmo de regresión y no solamente utilizar para comparar la regresión lineal. Además de buscar otros tipos de modelos con base en funciones con una sola variable independiente.

5. AGRADECIMIENTOS

Se agradece el apoyo de la Facultad de Ingeniería de la Universidad Autónoma de Chihuahua para la realización de éste trabajo.

6. REFERENCIAS

[1] Black C. L.; Merz C. J. UCI repository of machine learning databases. 1998.
[2] Blanco Vega, Ricardo; Blanco Vega, Humberto; Canales Leyva, Martha Guadalupe, Fernández Carrasco, Juan Carlos. *Metodología para el Desarrollo de Sistemas de Predicción***.** 1er. Congreso Internacional de Investigación (CIPITECH 2008). Cd. Juárez Chihuahua, México.
[3] Cleophas Ton J., Zwinderman Aeilko H. *Structural Equation Modeling (SEM) with SPSS Analysis of Moment Structures (Amos) for Cause Effect Relationships I (35 Patients).* Machine Learning in Medicine. 2015, pp 295-300
[4] Hasan Abid, Hasan Kamrul, Mottalib Abdul. *Linear regression–based feature selection for microarray data classification*. International Journal of Data Mining and Bioinformatics 2015 11:2, 167-179.
[5] Kim, M., & Pavlovic, V. *Dimensionality Reduction using Covariance Operator Inverse Regression*. IEEE. 2008.
[6] Kwak, N., Kim, C., & Kim, H. *Dimensionality reduction based on ICA for regression problems*. ScienceDirect. 2008.
[7] Marsland, S. *Machine Learning: an Algorithmic Perspective*. New York: Chapman & Hall/CRC. 2009.
[8] Molina Félix, Luis Carlos. Data mining: torturando a los datos hasta que confiesen. Coordinación del programa de Data mining Universidad Oberta de Catalunya. 2002. http://www.uoc.edu/molina1102/esp/art/molina1102/molina1102.html.
[9] Todorovski Ljupčo, Ljubič Peter, Džeroski Sašo. *Inducing Polynomial Equations for Regression. Machine Learning*: ECML 2004. pág. 441-452. Lecture Notes in Computer Science. Springer Berlin Heidelberg.
[10] Witten, I. H., & Frank, E. *Data mining*. New York: Morgan-Kaufmann. 2005.
[11] Xu, R.-F., & Lee, S.-J. *Dimensionality reduction by feature clustering for regression*. Science direct. 2015.