Regresión Local por Mínimos Cuadrados para Estimación Eficiente de Datos Incompletos

Pedro José García Laencina, José Luis Sancho Gómez
Departamento de Tecnologías de la Información y las Comunicaciones
Universidad Politécnica de Cartagena. Plaza del Hospital 1, 30202 Cartagena
Teléfono: (+34) 968326542. Fax: (+34) 968325973
E-mail: pedroj.garcia@upct.es

Resumen. *La presencia de valores perdidos o datos incompletos es un problema a solventar en muchas aplicaciones reales de reconocimiento de patrones. Un procedimiento extendido, y a la vez adecuado, es la imputación (i.e., estimación de valores perdidos a partir de la información conocida). Este artículo presenta un robusto algoritmo de imputación basado en la regresión local por mínimos cuadrados. Para cada patrón incompleto, se calculan sus K vecinos más cercanos, y a partir de esta información, la estimación de datos incompletos se obtiene mediante la resolución del problema de ajuste de mínimos cuadrados regularizado incluyendo el término de regularización de Tikhonov. Los resultados en un problema de diagnosis médica muestran las ventajas del método propuesto.*

1. Introducción

Un patrón[1] es una entidad que está representada por un conjunto de propiedades, conocido como vector de características [1]. Por ejemplo, en diagnosis médica, el vector de características está formado por los resultados de las distintas pruebas médicas realizadas en el paciente objeto de estudio. El objetivo en aprendizaje supervisado es entrenar modelos (máquinas de aprendizaje) [1] que predigan con exactitud nuevos valores de una tarea para futuras entradas. El término tarea se refiere a una función objetivo que es aprendida a partir de un conjunto de patrones (conocido como el conjunto de entrenamiento) [1]. Para implementar un sistema de ayuda a la diagnosis médica, podemos entrenar una máquina, a partir de una cierta cantidad de patrones representativos del problema a resolver, para que sea capaz de diagnosticar una determinada enfermedad en un nuevo paciente cuyos datos médicos no han sido utilizados durante el entrenamiento de la máquina.

La mayoría de bases de datos que caracterizan problemas reales tienen datos incompletos [2]. La ausencia de estos valores puede estar provocada por distintas causas, como de origen tecnológico (corte del suministro eléctrico) o porque simplemente son imposibles de medirse (un paciente no se puede someter a una prueba determinada). Una de las soluciones más empleadas es la imputación de datos [2], i.e., el proceso de estimar y rellenar datos incompletos a partir de toda la información disponible. En este trabajo, nos centramos en métodos de imputación basados en aproximaciones locales, donde destaca el algoritmo KNN (*K Nearest Neighbours*) como uno de los más extendidos. Este artículo propone una robusta implementación del método LLS (*Local Least Squares*) basada en la regularización de Tikhonov. El resto del artículo se estructura de la siguiente forma: la Sección 2 presenta la notación empleada y el algoritmo estandar KNN; la Sección 3 describe el método propuesto; la Sección 4 muestra los resultados obtenidos en un problema de diagnosis médica; finalmente, la Sección 5 expone las conclusiones principales.

*Este trabajo está financiado por el Ministerio de Educación y Ciencia a través del proyecto TEC2006-13338/TCM.

[1]Los términos patrón, vector de entrada, y caso son usados como sinónimos.

2. Métodos Locales de Imputación

Considerar un problema de aprendizaje supervisado caracterizado por una base de datos

$$\mathcal{D} = \{\mathbf{X}, \mathbf{M}, \mathbf{T}\} = \{\mathbf{x}_i, \mathbf{m}_i, t_i\}_{i=1}^{N}, \qquad (1)$$

donde $\mathbf{x}_i$ es el i-ésimo patrón compuesto por n características reales ($\mathbf{x}_i = \{x_{ij}\}_{j=1}^{n}$), $\mathbf{m}_i$ es un vector de variables binarias tal que m_{ij} es igual a 0 si x_{ij} está incompleto o 1 en caso contrario; y t_i es la salida deseada asociada a $\mathbf{x}_i$.

En lugar de usar el conjunto total de patrones, los métodos analizados en este trabajo están basados en aproximaciones locales, es decir, dado un patrón incompleto se obtiene un conjunto de los K vectores de entrada más similares (según una métrica de distancia) a dicho patrón, y se realiza una estimación de los valores perdidos usando la información disponible en ese conjunto de K casos similares. En concreto, la métrica escogida es la distancia euclidea. La distancia entre dos patrones $\mathbf{x}_a$ y $\mathbf{x}_b$ viene dada por la siguiente expresión,

$$d(\mathbf{x}_a, \mathbf{x}_b) = \sqrt{\sum_{j=1}^{n} (x_{aj} - x_{bj})^2 m_{aj} m_{bj}}. \qquad (2)$$

En el caso de que alguna(s) característica(s) en $\mathbf{x}_a$ y $\mathbf{x}_b$ sean desconocidas, dichas variables no son incluidas en $d(\mathbf{x}_a, \mathbf{x}_b)$.

Inicialmente, y para facilitar la descripción de los algoritmos, se asume que los patrones de entrada únicamente presentan un valor perdido en la primera característica, x_1. Además, $\mathbf{X}$ se divide en dos conjuntos: $\mathbf{X}^{\mathrm{C}}$ (casos completos) y $\mathbf{X}^{\mathrm{I}}$ (casos incompletos). Dado un patrón incompleto $\mathbf{x}$ con un valor perdido en x_1 (i.e., $x_1 = ?$ and $m_1 = 0$), sus K vecinos más cercanos (procedentes de $\mathbf{X}^{\mathrm{C}}$) son $\mathcal{V} = \{\mathbf{v}_k\}_{k=1}^{K}$, estando ordenados en orden creciente según $d(\mathbf{x}, \mathbf{v}_k)$. A partir de $\mathcal{V}$, se generan la *matriz de diseño* $\mathbf{A}$ y el vector $\mathbf{b}$. Así, $\mathbf{A} \in \mathbb{R}^{K\times(n-1)}$, donde sus K filas son los K vecinos más cercanos sin considerar sus valores en la primera característica (i.e., atributo incompleto a ser imputado). Además, $\mathbf{b} \in \mathbb{R}^{K\times 1}$ está compuesto por los valores de la primera característica de $\mathcal{V}$. Por último, $\mathbf{y} \in \mathbb{R}^{1\times(n-1)}$, cuyos elementos son los $(n-1)$ valores de los restantes atributos completos de $\mathbf{x}$. A continuación se muestra como $\mathbf{A}$, $\mathbf{b}$ e $\mathbf{y}$ son generados,

$$\begin{pmatrix} ? & \mathbf{y} \\ \mathbf{b} & \mathbf{A} \end{pmatrix} = \begin{pmatrix} ? & y_1 & \cdots & y_{n-1} \\ b_1 & a_{11} & \cdots & a_{1n-1} \\ \vdots & \vdots & \vdots & \vdots \\ b_k & a_{k1} & \cdots & a_{kn-1} \\ \vdots & \vdots & \vdots & \vdots \\ b_K & a_{K1} & \cdots & a_{Kn-1} \end{pmatrix} \quad (3)$$

$$= \begin{pmatrix} ? & x_2 & \cdots & x_n \\ v_{11} & v_{12} & \cdots & v_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ v_{k1} & v_{k2} & \cdots & v_{kn} \\ \vdots & \vdots & \vdots & \vdots \\ v_{K1} & v_{K2} & \cdots & v_{Kn} \end{pmatrix}. \quad (4)$$

2.1. Imputación KNN

La imputación obtenida mediante KNN viene dada por el promedio de los valores en la característica incompleta de los K vecinos más cercanos. Este método únicamente emplea la información relativa a la característica a imputar. Si $\mathbf{x}$ presenta un dato incompleto en x_1, el valor imputado por el procedimiento KNN es

$$\tilde{x}_1 = \frac{1}{K}\sum_{k=1}^{K} v_{k1} = \frac{1}{K}\sum_{k=1}^{K} b_k. \quad (5)$$

Una mejora directa se obtiene mediante una media ponderada,

$$\tilde{x}_1 = \frac{1}{K}\sum_{k=1}^{K} \alpha_k v_{k1} = \frac{1}{K}\sum_{k=1}^{K} \alpha_k b_k, \quad (6)$$

siendo $\alpha_k = \frac{\beta_k}{\sum_{k'=1}^{K}\beta_k}$, donde $\beta_k = \frac{1}{d(\mathbf{x},\mathbf{v}_k)^2}$.

3. Método Propuesto

3.1. Imputación LLS

El algoritmo LLS utiliza toda la información local en $\mathcal{V}$. Para ello se plantea el siguiente problema de regresión por mínimos cuadrados [3]

$$\min_{\mathbf{z}} \|\mathbf{b} - \mathbf{A}\mathbf{z}\|^2, \quad (7)$$

siendo la solución exacta a dicho problema el vector columna

$$\hat{\mathbf{z}} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b}. \quad (8)$$

Así, el dato incompleto puede ser obtenido mediante la combinación lineal

$$\tilde{x}_1 = \mathbf{y}\hat{\mathbf{z}}. \quad (9)$$

Este método puede directamente extenderse al problema general con más de un dato incompleto, i.e., $\mathbf{x}$ presenta Q valores perdidos, con $Q > 1$. Para ello se genera $\mathbf{B} \in \mathbb{R}^{K\times Q}$, donde cada vector columna está formado por los valores de las q-ésimas características incompletas $(1 \leq q \leq Q)$ de los K vecinos más cercanos, y por tanto ahora $\mathbf{A} \in \mathbb{R}^{K\times(n-Q)}$. Además, $\mathbf{y}$ está compuesto por los $n-Q$ valores observados (características completas) en $\mathbf{x}$. Tras generar $\mathbf{A}$, $\mathbf{B}$ e $\mathbf{y}$, se obtiene la solución $\hat{\mathbf{z}}$ de $\min_{\mathbf{z}} \|\mathbf{B} - \mathbf{A}\mathbf{z}\|^2$, y los Q valores perdidos en $\mathbf{x}$ son estimados por $\mathbf{y}\hat{\mathbf{z}}$.

3.2. Término de Regularización

La solución exacta al problema de mínimos cuadrados puede producir una deficiente estimación de valores perdidos cuando $\mathbf{A}$ no es de rango completo o $\mathbf{A}$ está mal condicionada [3]. Para obtener una solución estable, este artículo emplea el método de regularización de Tikhonov [4]. De esta forma, el problema regularizado viene dado por

$$\min_{\mathbf{z}} \|\mathbf{b} - \mathbf{A}\mathbf{z}\|^2 + \lambda\|\mathbf{z}\|^2. \quad (10)$$

El primer término representa el error de predicción, mientras que el segundo término añade información a priori de la solución, penalizando un valor elevado de la norma del vector solución $\mathbf{z}$. Para un determinado valor del parámetro de regularización λ, la solución regularizada es

$$\hat{\mathbf{z}}_\lambda = (\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b}. \quad (11)$$

donde $\mathbf{I}$ es la matriz identidad. El método lo denominamos rLLS (regularized LLS). Para $\lambda = 0$, la solución $\hat{\mathbf{z}}_\lambda$ se reduce al problema original. Si $\lambda \to \infty$, $\hat{\mathbf{z}}_\lambda$ tiende a cero para minimizar la norma de la solución. El valor adecuado de λ combinará del mejor modo ambas soluciones. Además para cualquier valor de λ es posible obtener $(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda\mathbf{I})^{-1}$, incluso si $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$ no existe. El valor óptimo de λ se escoge mediante el método de validación cruzada generalizada [5], que es un conocido y eficiente procedimiento para obtener λ.

4. Resultados Experimentales

Con el objetivo de comparar el método propuesto con la imputación KNN ponderada, se ha empleado un problema de diagnosis de cáncer de mama, *Wisconsin Diagnostic Breast Cancer* (WDBC) [6]. El conjunto de datos consta de 569 casos, compuestos por 30 variables de entrada. Tras normalizar los datos de entrada con media cero y varianza unidad, se eliminan aleatoriamente distintos porcentajes de datos ($5\,\%, 10\,\%, 20\,\%, 40\,\%, 60\,\%$) en la característica x_{24} (relevante para la diagnosis) para evaluar los métodos KNN y rLLS (con $K = [5, 10, 15, 20]$).

A continuación se presentan los resultados de imputación para un $60\,\%$ de datos incompletos en x_{24} y para 15 vecinos más cercanos. La Figura 1 muestra las funciones de distribución empíricas en x_{24}, considerando los datos originales y los datos imputados con KNN y rLLS.

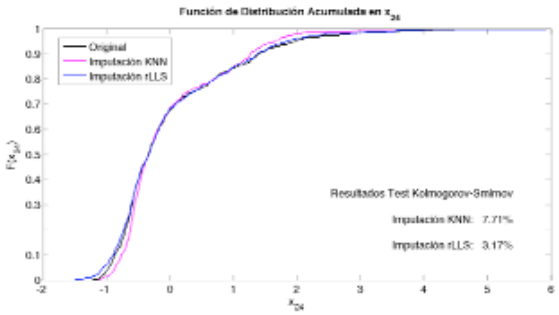

Fig. 1. Funciones de distribución empíricas original y tras imputación mediante KNN y rLLS, con un $60\,\%$ de datos incompletos en x_{24} y $K = 15$.

Como se puede ver en dicha figura, el método propuesto proporciona una mejor solución que KNN, manteniendo la función de distribución original. Para evaluar las prestaciones en términos de la distribución (*Distributional Accuracy*, DAC) hemos usado el test de Kolmogorov-Smirnov, que calcula la mayor diferencia entre la distribución original y la distribución de los datos imputados.

La Figura 2 muestra los diagramas de dispersión (*scatter diagram*) para la característica x_{24}, representando los datos originales en el eje de abcisas y los datos imputados mediante KNN y rLLS en el eje de ordenadas.

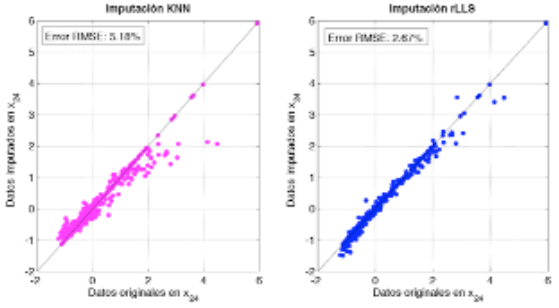

Fig. 2. Datos originales vs. Datos imputados mediante KNN y rLLS, con un $60\,\%$ de datos incompletos en x_{24} y $K = 15$.

Además muestra la función identidad, ya que dicha relación lineal existe cuando las imputaciones coinciden con los valores originales. Podemos comprobar como las estimaciones obtenidas por el método rLLS se aproximan con mayor precisión a los valores reales que en el caso del método KNN. Para evaluar las prestaciones en términos de la predicción (*Predictive Accuracy*, PAC) se emplea la raiz del error cuadratico medio (*Root Mean Square Error*, RMSE).

Por último, las Tablas I y II muestran los resultados obtenidos (DAC y PAC, respectivamente) con los métodos de imputación KNN y rLLS para distintos porcentajes de datos incompletos y valores de K. El método rLLS proporciona mejores prestaciones que el algoritmo KNN en todos los experimentos realizados. Cabe destacar que en este problema los resultados de la imputación KNN degradan conforme aumenta K, al contrario que en el método propuesto.

DAC (%)		% de valores perdidos en x_{24}				
		5	10	20	40	60
$K = 5$	KNN	0.80	1.30	2.12	3.73	5.97
	rLLS	0.68	0.89	1.35	2.39	2.81
$K = 10$	KNN	0.86	1.54	2.38	4.43	7.14
	rLLS	0.65	0.88	1.42	2.32	3.43
$K = 15$	KNN	0.88	1.66	2.58	4.76	7.71
	rLLS	0.64	0.86	1.40	2.24	3.17
$K = 20$	KNN	0.89	1.68	2.73	4.99	8.26
	rLLS	0.52	0.84	1.24	2.14	2.59

TABLA I

DAC (%) OBTENIDA MEDIANTE KNN Y RLLS.

PAC (%)		% de valores perdidos en x_{24}				
		5	10	20	40	60
$K = 5$	KNN	3.98	4.31	4.24	4.15	4.71
	rLLS	2.85	2.98	3.32	3.25	3.64
$K = 10$	KNN	3.99	4.46	4.52	4.49	4.89
	rLLS	2.13	2.38	2.77	2.78	2.98
$K = 15$	KNN	4.07	4.75	4.84	4.77	5.18
	rLLS	1.87	2.19	2.57	2.48	2.67
$K = 20$	KNN	4.30	5.02	5.10	5.02	5.45
	rLLS	1.63	1.84	2.20	2.23	2.23

TABLA II

PAC (%) OBTENIDA MEDIANTE KNN Y RLLS.

5. Conclusiones

Este artículo presenta una versión eficiente y robusta del algoritmo LLS para estimación de datos incompletos. El método propuesto está basado en el uso del término de regularización de Tikhonov. Los resultados obtenidos muestran las ventajas del método propuesto sobre el algoritmo KNN.

Referencias

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, New York, USA, 2006.

[2] R. J. A. Little, D. B. Rubin, *Statistical Analysis with Missing Data*. Wiley, New Jersey, USA, 2nd edition, 2002.

[3] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, 1996.

[4] A. N. Tikhonov, Solution of incorrectly formulated problems and the regularization method. *Sov Math Dokl*, pp. 1035-1038, vol. 4, 1963.

[5] G. H. Golub, U. Matt, "Generalized cross-validation for large-scale problems". *J Comput Graph Stat*, pp. 1-34, vol. 6, no. 1, 1997.

[6] http://www.ics.uci.edu/~mlearn/MLRepository.html