Extendiendo la Semántica de los datos en Aprendizaje Supervisado

Article · October 2003

CITATIONS
0

READS
50

5 authors, including:

Gualberto Asencio Cortés
Universidad Pablo de Olavide
42 PUBLICATIONS **229** CITATIONS
SEE PROFILE

José C Riquelme
Universidad de Sevilla
256 PUBLICATIONS **2,648** CITATIONS
SEE PROFILE

Jesús S. Aguilar-Ruiz
Universidad Pablo de Olavide
227 PUBLICATIONS **2,717** CITATIONS
SEE PROFILE

Francisco Javier Ferrer-Troyano
Universidad de Sevilla
32 PUBLICATIONS **152** CITATIONS
SEE PROFILE

Extendiendo la Semántica de los datos en Aprendizaje Supervisado

Gualberto Asencio, José C. Riquelme, Jesús S. Aguilar-Ruiz, Francisco J. Ferrer
Departamento de Lenguajes y Sistemas Informáticos. Universidad de Sevilla
gualberto@arrakis.es, {riquelme,aguilar,ferrer}@lsi.us.es

Resumen. En este trabajo, centrado en el área del aprendizaje supervisado, pretendemos extender la información que proporcionan los datos etiquetados. Basándonos en la técnica de los vecinos más cercanos, se amplía la información contenida en las etiquetas discretas de las instancias, fortaleciendo su semántica y perfeccionando una clasificación posterior. Para ello, por cada instancia del conjunto de entrenamiento se obtiene información acerca de las distancias a las instancias más cercanas de clase contraria. Mediante un modelo geométrico se transforma cada etiqueta y es extendida y representada por un punto en $\Re^{n-1}$ (donde n es el número de valores distintos de la clase discreta). Utilizando el conjunto de datos de UCI, hemos realizado dos experimentos para validar nuestra propuesta y verificar que no existe pérdida de información y que es posible nuevas prestaciones al reemplazar una clase discreta por una continua.

1. Introducción

En la clasificación de instancias por vecinos más cercanos (k-NN) [2], dada una instancia a clasificar se obtienen las k instancias (vecinos) que aplicándoles una función de distancia determinada (generalmente Euclídea) son más cercanas a la instancia a clasificar. A continuación se clasifica la nueva instancia asignándole la clase mayoritaria (si es discreta) o una media ponderada (si es continua) de entre sus vecinos anteriormente calculados.

Normalmente una clasificación para una etiqueta o clase continua proporciona algunas facilidades que no se tienen con las etiquetas discretas. La principal es que la función de error para los ejemplos del fichero de entrenamiento es una función real y continua: por ejemplo, la suma de las diferencias en valor absoluto entre el valor estimado y el valor conocido. Con una etiqueta discreta la función de error contaría el número de ejemplos mal clasificados, que es una función entera y escalonada. La ventaja de minimizar una función continua es innegable: pequeñas variaciones en los parámetros que definen el modelo de clasificación dan lugar a pequeñas variaciones en la función de error, lo que permite la aplicación de técnicas de búsqueda de óptimos más eficientes.

Nuestro objetivo es extender la semántica o significado de una etiqueta discreta mediante su transformación en continua utilizando para ello el modelo de clasificación del vecino más cercano. De esta forma, habremos recuperado las interesantes propiedades antes mencionadas para etiqueta continua, con la ventaja de que se añade nueva información que robustece el significado de la etiqueta discreta.

Hemos diseñado un modelo geométrico al que trasladamos las instancias del conjunto de entrenamiento. Sustituimos las etiquetas de cada instancia por un punto

en $\Re^{n-1}$, donde n es el número de valores distintos de la etiqueta discreta. Este punto se asignará en función del valor de su clase y de las distancias a sus enemigos más cercanos. Los enemigos más cercanos son las instancias de clase distinta que están más cerca. Hay por consiguiente n-1 enemigos más cercanos. A la hora de clasificar, promediar los vecinos más cercanos consiste en promediar los puntos de $\Re^{n-1}$ calculados para dichos vecinos.

Como veremos con detalle, la traslación de las instancias a este modelo geométrico supone una carga de información valiosa a las instancias no inherente en sus clases y robustece su semántica perfeccionando las clasificaciones.

Para validar el modelo se ha utilizado un clasificador basado en el vecino más cercano sobre clase continua en la herramienta Weka [3]. En el apartado 5 mostramos las pruebas realizadas con el mismo sobre el almacén UCI [1] y comparamos resultados con un clasificador k-NN convencional. Por último, comprobamos cómo la extensión de esta semántica permite añadir información sobre probabilidad de acierto en clasificadores que a priori no la tienen.

2. Modelo geométrico

La transformación de las etiquetas discretas de las instancias del conjunto de entrenamiento en puntos de $\Re^{n-1}$ se hace en función del valor de su clase y de las distancias a sus enemigos más cercanos. La ubicación de los puntos en $\Re^{n-1}$ obtenidos para las instancias supone una transición suave entre los diferentes valores para la etiqueta discreta. La construcción del modelo geométrico que extiende la semántica de las instancias la abordaremos detalladamente en el siguiente apartado. Ahora vamos a definir los conceptos básicos sobre los que se apoya.

Se fijan unos puntos en $\Re^{n-1}$ llamados puntos base B_i correspondientes a los distintos valores para la etiqueta discreta. Habrá por tanto n puntos base: B_1, …, B_n. Estos puntos están separados a igual distancia entre sí, por ejemplo, distancia unitaria. Todas las etiquetas obtenidas a partir de las etiquetas discretas de las instancias son puntos que pertenecen a la superficie (n-1)-dimensional formada por los puntos base.

Así, por ejemplo, si se tratase de unos datos con una clase discreta con tres valores distintos, habría tres puntos base formando un triángulo equilátero en $\Re^2$. Todos los puntos que representarán las nuevas etiquetas de las instancias, al transformarlas, pertenecen a dicho triángulo. Por tanto, y como se ha considerado distancia unitaria entre los puntos base, todas las distancias entre los puntos que consideremos estarán comprendidas entre 0 y 1.

Nuestro objetivo es proporcionar a cada instancia una nueva etiqueta (un punto F en $\Re^{n-1}$) que sustituya la original y proporcione información añadida conservando ésta. La distancia de F a los puntos base determina que dicha instancia sea de una clase u otra. De esta forma, si una instancia es de clase A y el punto base B_1 representa a la clase A, el punto F asociado a la instancia estará más cerca de B_1 que de cualquier otro punto base.

Ejemplo 1.

Sea un fichero de entrenamiento con instancias etiquetadas con tres clases posibles A, B y C. Después de un preprocesamiento que se detalla en el siguiente apartado,

este fichero queda constituido por las mismas instancias donde las etiquetas discretas han sido sustituidas por puntos en $\Re^2$, que constituyen una nueva etiqueta "continua". Se pretende clasificar una instancia Q. Puesto que son tres valores distintos para la etiqueta discreta, trabajaremos por tanto en $\Re^2$. Sean los puntos base: B_1=(0,0), B_2=(1,0) y B_3=$(0.5, \sqrt{3}/2)$. Estos puntos (ver Fig. 1) constituyen un triangulo equilátero de base el intervalo [0,1] y representan respectivamente las etiquetas A, B y C. En el apartado 3 veremos el proceso de creación de dichos puntos base y de los puntos que extienden las etiquetas de las instancias que, en este ejemplo, denominamos F_1, F_2 y F_3.

Sea el número de vecinos más cercanos con el que se clasifica, k = 3. Supongamos calculados dichos vecinos cuyas etiquetas son: A, A y C. Nótese que un clasificador por vecinos más cercanos convencional hubiese clasificado la instancia Q como de clase A ya que hay mayoría de clase A en los vecinos más cercanos. El proceso que llevaríamos a cabo para clasificar con nuestro modelo es el siguiente:

Vecino 1:

- Es de clase A
- Nueva etiqueta extendida F_1=(0.35, 0.25)
  (En el siguiente apartado se explicará cómo se obtiene este valor)
- Distancia de F_1 a B_1 = 0.430; de F_1 a B_2 = 0.696; de F_1 a B_3 = 0.634;
- Observamos que el punto F_1 está más cerca de B_1, el punto base asociado a la clase A, que de B_2 y B_3. No obstante lo está por poco, nótese que la distancia a B_1 no es mucho menor que las otras.

Vecino 2:

- Es de clase A
- Nueva etiqueta extendida F_2=(0.42,0.15)
- Distancia de F_2 a B_1 = 0.446; de F_2 a B_2 = 0.599; de F_2 a B_3 = 0.720;
- Observamos que el punto F_2 está más cerca de B_1, el punto base asociado a la clase A, que de B_2 y B_3. En este caso ocurre como en el punto anterior, la distancia a B_1 no es mucho menor que las otras.

Vecino 3:

- Es de clase C
- Nueva etiqueta extendida F_3=(0.5,0.8)
- Distancia de F_3 a B_1 = 0.943; de F_3 a B_2 = 0.943; de F_3 a B_3 = 0.066;
- Observamos que el punto F_3 está más cerca de B_3, el punto base asociado a la clase C. En este caso lo está de forma acentuada puesto que la distancia a B_3 es bastante menor que las otras.

Clasificación de Q mediante el promedio de las nuevas etiquetas de sus vecinos:

- El punto promedio $Z = \frac{F_1 + F_2 + F_3}{k} = \frac{(0.35,0.25) + (0.42,0.15) + (0.5,0.8)}{3} = (0.42333,0.4)$
- Distancia de Z a B_1 = 0.582; de Z a B_2 = 0.702; de Z a B_3 = 0.472;
- Observamos que el punto Z está más cerca de B_3, el punto base asociado a la clase C. La diferencia en este caso no es muy acusada, por tanto, podemos clasificar Q como de clase C e incluso asignar una probabilidad de éxito a la clasificación, por ejemplo, el complementario de la distancias normalizadas a dos (para que las probabilidades sumen 1). Es decir, si sumamos las distancias de Z a los puntos bases obtenemos 1.756, realizando una normalización a [0,2] (2

porque estamos en R^2), las distancias serían respectivamente 0.66 (= 2 × 0.582 / 1.756), 0.80 (= 2 × 0.702 / 1.756) y 0.54 (= 2 × 0.472 / 1.756) a cada clase. Luego podríamos afirmar que Q sería de clase A con probabilidad 0.33, de clase B con probabilidad 0.2 y de clase C con probabilidad 0.46.

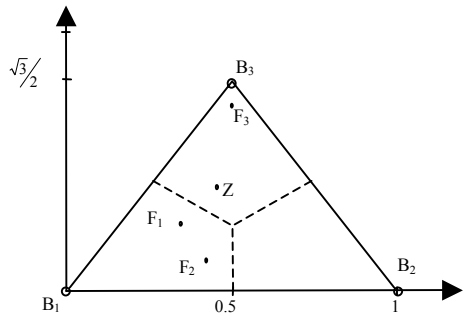

Fig. 1 Representación gráfica del ejemplo 1

La construcción del modelo geométrico ha permitido decantarse por la clase C en la clasificación. Esto es porque los vecinos de clase A (el 1º y el 2º) están muy cerca de instancias de clase contraria (enemigos). En la traducción de dichos vecinos al modelo que nos ocupa, se observa en la figura 1 que resultan puntos cercanos a la frontera de la región de Voronoi de B_1, punto base asociado a la etiqueta A.

Ocurre todo lo contrario con el vecino 3º que es de clase C y lejano a instancias de clase contraria. Por eso está notablemente más cerca de B_3 que de los puntos base B_1 y B_2. Finalmente, como se observa en la figura 1, el punto promedio Z de F_1, F_2 y F_3 se encuentra en la región de Voronoi de B_3 por lo que la clase ganadora ha sido C.

De este modo, en la clasificación de instancias con etiqueta discreta se ha tenido en cuenta no solo los valores de la etiqueta de los vecinos más cercanos sino su proximidad relativa a instancias de clase contraria lo que hace más rica la clasificación en cuanto a cuestiones de semántica se refiere.

3. Extendiendo la semántica

Como hemos expuesto y ejemplificado en el apartado anterior, el modo que empleamos para extender la semántica de las instancias es la creación de un modelo geométrico al que trasladamos las mismas. Posteriormente, durante la clasificación de una instancia, promediamos los puntos obtenidos de los vecinos más cercanos y observamos la distancia de dicho punto promedio a los puntos base. Clasificaremos con la etiqueta asociada al punto base del que menos diste.

Como se pudo apreciar en el ejemplo 1, los puntos que representan las nuevas etiquetas asociadas a las instancias recogen información de la clase de la instancia. También llevan información no sólo de la distancia a sus enemigos más cercanos, sino a qué enemigos (uno por cada clase distinta a la de la instancia) y a qué distancia para cada uno de ellos. Antes de reetiquetar las instancias, necesitamos crear el

modelo geométrico en el que situar las instancias del conjunto de entrenamiento. Describiremos en el punto 3.1 el proceso de creación del modelo geométrico y en el punto 3.2, cómo para cada etiqueta de cada instancia obtenemos su punto en el modelo. Una vez obtenidos todos los puntos para todas las instancias del conjunto de entrenamiento, el preprocesado habrá terminado y se podrán hacer clasificaciones como se hizo en el ejemplo 1.

3.1 Construcción del modelo

Para simplificar la notación, a los valores de la etiqueta discreta le asociamos un número natural. De modo que si son por ejemplo: A, B y C, le asociaremos los valores 1, 2 y 3. Para construir el modelo realizamos los siguientes pasos:

3.1.1 Construcción de los puntos base B_i

Sea n, como utilizamos anteriormente, el número de valores distintos de la etiqueta discreta. Hay por tanto que construir n puntos en $\Re^{n-1}$ separados entre sí a distancia unitaria. Utilizamos $\Re^{n-1}$ porque es el espacio vectorial de menor dimensión en el que existen n puntos separados a igual distancia. Como apuntamos anteriormente, los puntos base los notaremos con $B_i \in \Re^{n-1}$, $i = 1...n$. Como es natural, n>=2 y por ello fijamos dos puntos base iniciales: B_1=(0, …, 0) y B_2=(1, 0, …, 0).

En la figura 2 izquierda se encuentra la representación de 3 puntos base B_1, B_2 y B_3 en $\Re^2$ formando un triángulo equilátero. Las líneas discontinuas representan los límites de las regiones de Voronoi de los puntos base a partir de los cuales las instancias se clasifican con una etiqueta u otra.

En la figura 2 derecha se ha representado los 4 puntos base en $\Re^3$ formando un tetraedro regular. Las líneas discontinuas representan los límites de la región de Voronoi de B_1.

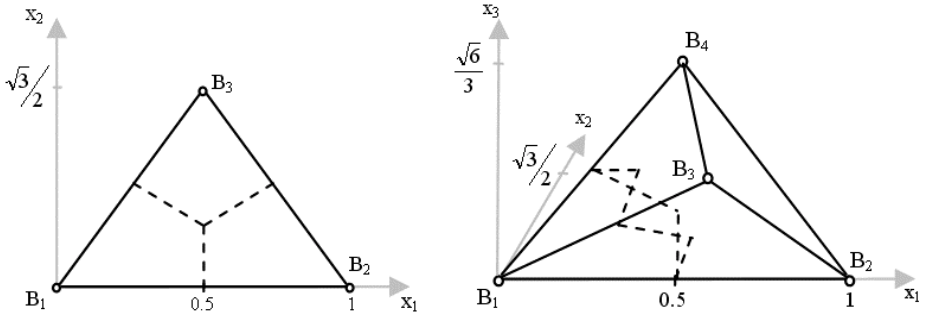

Fig. 2. Disposición de los puntos base para tres etiquetas (izq.) y cuatro etiquetas (der.)

El proceso para construir los puntos base es un proceso recursivo puesto que para calcular el i-ésimo punto es necesario imponer que se encuentre a distancia unitaria de los demás lo cual implica que estos deben haber sido calculados previamente. De esta forma, este proceso está expresado para calcular el i-ésimo punto base y supone calculados los i-1 puntos base anteriores.

1. Dados los puntos base B_1, …, B_{i-1}, obtenemos el baricentro V de los mismos.

$$V = \frac{\sum_{j=1}^{i-1} B_j}{i-1} \in \Re^{n-1} \quad (\mathbf{1})$$

El baricentro se encuentra en la superficie formada por los puntos base B_1,.., B_{i-1}, y estos, por construcción, tienen sus n-i últimas componentes a 0. Esto es así puesto que este proceso añade una y sola una componente distinta de 0 cada vez. Por consiguiente, se tiene que $V_j = 0$ para $i \le j \le n-1$.

2. Obtenemos la ecuación de la recta r normal a la superficie (i-1)-dimensional formada por B_1, …, B_{i-1} que pasa por el punto V. La ecuación de la recta r se expresa fijando las i-2 primeras coordenadas y dejando libre x_{i-1}:

$$r: x_j = V_j \text{ para } 1 \le j < i-1 \text{ y } x_{i-1} = \lambda \in \Re \quad (\mathbf{2})$$

3. Calculamos la intersección de la recta r con el lugar geométrico de los puntos que distan 1 desde cualquier punto base ya calculado. Esta intersección es el nuevo punto base B_i. Sea B_{ij} con j= 1...n-1 las coordenadas del punto B_i. Su cálculo es el siguiente: Para $1 \le j < i-1$, $B_{ij} = V_j$ y para $i \le j \le n-1$, $B_{ij} = 0$. Para obtener $B_{i,i-1}$, tenemos que $x_1^2 + x_2^2 + ... + x_{i-1}^2 = 1$ es el lugar geométrico antes mencionado tomando como origen B_1. Imponemos que $x_1 = V_1, x_2 = V_2, ..., x_{i-2} = V_{i-2}$ y obtenemos que $x_{i-1} = B_{i,i-1}$. Despejando $B_{i,i-1}$ resulta:

$$B_{i,i-1} = \sqrt{1 - \sum_{j=1}^{i-2} V_j^2} \quad (\mathbf{3})$$

Ejemplo 2.

Supongamos que para unos datos, la etiqueta o clase discreta tiene 4 valores diferentes numerados del 1 al 4. Por tanto, hay que construir 4 puntos base: B_1,..,B_4 en $\Re^3$ separados entre sí a distancia unitaria.

- $B_1 = (0, 0, 0)$, $B_2 = (1, 0, 0)$
- Cálculo de B_3: – Baricentro $V = \frac{B_1 + B_2}{2} = \frac{(0,0,0)+(1,0,0)}{2} = (0.5,0,0)$
  - Ecuación de la recta normal r: $x_1 = V_1 = 0.5$; $x_2 = \lambda \in \Re$
  - Intersección de la recta r con la circunferencia de radio 1 centrada en B_1:
    - Ecuación de dicha circunferencia: $x_1^2 + x_2^2 = 1$
    - Imponemos que x_1=0.5 y despejamos x_2
    - $x_2 = \sqrt{1 - x_1^2} = \sqrt{1 - 0.5^2} = \frac{\sqrt{3}}{2}$
  - Finalmente, $B_3 = (V_1, x_2, 0) = (0.5, \sqrt{3}/2, 0)$
- Cálculo de B_4:
  - Baricentro $V = \frac{B_1 + B_2 + B_3}{3} = \frac{(0,0,0) + (1,0,0) + (0.5, \sqrt{3}/2, 0)}{3} = (0.5, \frac{\sqrt{3}}{6}, 0)$

- Ecuación de la recta normal r: $x_1 = V_1 = 0.5$; $x_2 = V_2 = \sqrt{3}/6$; $x_3 = \lambda \in \Re$
- Intersección de la recta r con la esfera de radio 1 centrada en B_1:
  - Ecuación de dicha esfera: $x_1^2 + x_2^2 + x_3^2 = 1$
  - Imponemos que x_1=0.5 y $x_2 = \sqrt{3}/6$ y despejamos x_3
  - $x_3 = \sqrt{1 - x_1^2 - x_2^2} = \sqrt{1 - 0.5^2 - \left(\sqrt{3}/6\right)^2} = \frac{\sqrt{6}}{3}$
- Finalmente, $B_4 = (V_1, V_2, x_3) = (0.5, \sqrt{3}/6, \sqrt{6}/3)$

3.1.2 Obtención de los puntos medios M_{ij}

Ahora vamos a hallar los puntos medios de los segmentos que unen los puntos base. Como veremos en breve, estos puntos son necesarios para poder construir el modelo y enfrentar los diferentes valores de la etiqueta situados en los puntos base.

Hay $\frac{n(n-1)}{2}$ puntos medios los cuales denotaremos por $M_{i,j}$ donde i y j son los índices de los puntos base enfrentados. Es decir, $M_{1,2}$=$M_{2,1}$ es el punto medio entre B_1 y B_2. De esta forma,

$$M_{i,j} = M_{j,i} = \frac{B_i + B_j}{2}, \forall i, j = 1...n, i \neq j \quad (4)$$

3.2 Extensión de la semántica de las instancias

Los pasos descritos en 3.1 son independientes de instancias y se pueden calcular una sola vez. Para el proceso de extensión de la semántica que describimos en este punto se considera fijada una instancia T del conjunto de entrenamiento. Supongamos que dicha instancia es de clase i. Para llevar a cabo esta extensión realizamos los pasos que se describen a continuación.

3.2.1 Construcción de los puntos H_j

Para considerar las distancias a los enemigos más cercanos de la instancia T e incluir esa información en nuestro modelo, vamos a crear unos puntos (uno por cada etiqueta diferente) que llamaremos H_j $(1 \leq j \leq n \text{ e } i \neq j)$ situados en los segmentos que tienen origen en B_i (i es la clase de la instancia T) y extremo en los puntos medios M_{ij}. Llamaremos ned (nearest enemy distance) al vector de $\Re^n$ que contiene las distancias a los enemigos más cercanos para la instancia T. De esta forma, ned_j es la distancia entre T y la instancia de clase j más cercana a T con $i \neq j$. La situación de H_j dentro del segmento $\overline{B_i M_{ij}}$ está determinada por ned_j de esta forma:

$$H_j = ned_j \cdot B_i + (1 - ned_j) \cdot M_{ij} \quad (5)$$

Como se puede apreciar en la figura 3, los puntos H_j suponen una disposición en el espacio vectorial $\Re^{n-1}$ que permite enfrentar el punto base i con los restantes

puntos base en función de la distancia desde la instancia T a las más cercanas de clases contrarias. Como se verá a continuación, estos puntos servirán para construir el punto final que sustituye a la etiqueta discreta de la instancia T.

3.2.2 Obtención de las contribuciones finales α_j

Consideramos la superficie S_H (n-1)-dimensional formada por los puntos H_j. El punto final F que extiende la etiqueta de la instancia T y la sustituye en la clasificación es un punto que pertenece a S_H. La ubicación de F dentro de S_H viene determinada por unos coeficientes que llamaremos α_j. Estos están definidos del siguiente modo:

$$\alpha_j = \frac{\text{ned}_j}{\sum_{\substack{c=1 \\ c\neq i}}^{n} \text{ned}_c} \in [0,1] \qquad \sum_{\substack{j=1 \\ j\neq i}}^{n} \alpha_j = 1 \tag{6}$$

3.2.3 Obtención del punto final F

Finalmente, el punto F, que conforma la nueva etiqueta, se obtiene como el promedio ponderado por las contribuciones α_j de los puntos H_j.

$$F = \sum_{\substack{j=1 \\ i\neq j}}^{n} \alpha_j \cdot H_j \tag{7}$$

En la figura 3 correspondiente al ejemplo 3 se muestra una representación de los elementos construidos en los apartados anteriores.

Ejemplo 3.

Dados unos datos de entrenamiento, se desea reetiquetar las instancias extendiendo su semántica. Las instancias son de clase discreta con valores: A, B y C. Ejemplificaremos el proceso seguido para una instancia en concreto del conjunto de entrenamiento. Sea A la clase de dicha etiqueta y las distancias mínimas a instancias de otras clases son 0.75 para clase B y 0.3 para clase C que llamaremos respectivamente ned_2 y ned_3. Dado que existen tres valores diferentes para la clase discreta, empleamos los 3 puntos base: B_1, B_2 y B_3 utilizados en el ejemplo 1 y calculados en el ejemplo 2.

En primer lugar obtenemos los 3(3-1) / 2 = 3 puntos medios:

- $M_{12} = M_{21} = \frac{B_1 + B_2}{2} = (0.5,0)$; $M_{13} = M_{31} = (0.25, 0.43)$; $M_{23} = M_{32} = (0.75,0.43)$

A continuación calculamos los puntos H_j con j = 2, 3:

- $H_2 = \text{ned}_2 \cdot B_1 + (1 - \text{ned}_2) \cdot M_{12} = 0.75 \cdot (0, 0) + 0.25 \cdot (0.5, 0) = (0.125, 0)$
- $H_3 = \text{ned}_3 \cdot B_1 + (1 - \text{ned}_3) \cdot M_{13} = 0.3 \cdot (0, 0) + 0.7 \cdot (0.25, 0.43) = (0.175, 0.301)$

Una vez calculados los puntos H_j, obtenemos las contribuciones finales α_j:

- $\alpha_2 = \frac{ned_2}{\sum_{c=2}^{3} ned_c} = \frac{0.75}{0.75 + 0.3} = 0.7143 \quad \alpha_3 = \frac{ned_3}{\sum_{c=2}^{3} ned_c} = \frac{0.3}{0.75 + 0.3} = 0.2857$

Finalmente el punto F resultante es:

- $F = \sum_{j=2}^{3} \alpha_j \cdot H_j = ned_2 \cdot H_2 + ned_3 \cdot H_3 = 0.75 \cdot (0.125,0) + 0.3 \cdot (0.175,0.3) = (0.146,0.09)$

Como puede apreciarse en la figura 3, este punto F pertenece a la región de Voronoi del punto B_1 (correspondiente a la clase A) y su ubicación en la misma informa de las distancias a instancias de clase contraria.

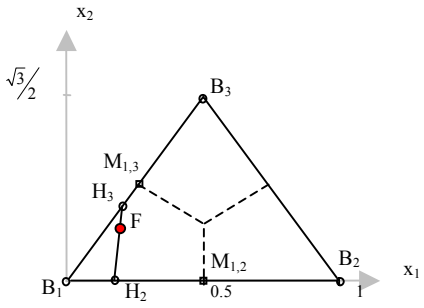

Fig. 3. Obtención de la nueva etiqueta F

4. Pruebas realizadas

Para probar que la extensión realizada no hace perder información a los datos originales se ha aplicado nuestro modelo a un subconjunto de bases de datos del almacén UCI. Una vez realizada la reetiquetación se ejecutó un clasificador kNN (denominado xNN) para clase continua. En la siguiente tabla se exponen los promedios de los porcentajes de aciertos que resultan de ejecutar el clasificador xNN para los valores de k: 1, 3, 5 y 7. Para ello se ha utilizado validación cruzada *leaving one out*, para que los resultados no dependan de la elección de las bolsas. Se enfrentan estos resultados con los de un clasificador por vecinos más cercanos clásico: kNN para los datos originales, es decir, con la etiqueta discreta.

Como se puede apreciar en la tabla 1, el clasificador para etiqueta continua prácticamente iguala (el test t de diferencia de medias no es significativo en ningún caso) al clasificador convencional por vecinos más cercanos kNN. Con esto se demuestra que la nueva etiqueta conserva la información que tenía la etiqueta discreta.

Otra evaluación realizada consiste en explotar la potencialidad que representa una clase continua frente a una etiqueta discreta. Por ejemplo, proporcionando información añadida sobre probabilidad de clasificación correcta. Existen problemas de clasificación donde una respuesta errónea representa un coste excesivo (pensemos en datos médicos).

Con los clasificadores de etiqueta discreta habituales es difícil proporcionar información sobre si para un punto es o no más probable una clasificación correcta.

BD	PromedioAciertos								
	k = 1		k = 3		k = 5		k = 7		
	kNN	xNN	kNN	xNN	kNN	xNN	kNN	xNN	
audiology	79,2	79,2	68,1	69,5	61,9	62,4	57,1	56,6	=++-
autos	74,6	74,6	67,8	70,2	63,9	69,8	59,5	60,5	=+++
balance-sc	88,0	88,6	88,0	88,6	88,3	89	89,6	90,2	++++
cleveland-1	75,2	75,2	82,5	81,8	82,2	80,9	82,5	82,2	=---
german_cr	72,9	72,9	72,6	72,8	73,6	73,4	73,1	73,3	=+-+
Glass	70,1	70,1	70,1	71,5	66,8	69,2	63,1	66,8	=+++
lymphogra	82,4	81,8	81,8	81,8	85,1	85,1	83,8	84,5	-==+
pima_diab	70,6	70,6	74,1	74	74,1	74,3	75,3	75,1	=-+-
primary-tur	39,8	39,5	44,5	43,4	47,2	46	47,8	47,5	----
sonar	87,5	87,5	83,7	83,7	82,2	82,2	80,3	80,8	===+
soybean	91,4	91,4	91,4	92,1	90,2	90,3	89,3	89,1	=++-
vehicle	69,7	69,7	70,4	70,6	70,3	70,5	70,9	71,1	=+++
vote	92,4	92,9	92,9	93,3	93,1	93,6	93,1	93,6	++++
vowel	99,4	99,4	98,0	98,0	95,3	95,3	90,3	90,6	===+
zoo	96,0	97	92,1	91,1	95,0	95	90,1	91,1	+-=+

Tabla 1. Resultados obtenidos para xNN y kNN en las bases de datos de UCI

Con las etiquetas continuas definidas en este trabajo, es posible proporcionar una noción de cómo de lejos estamos de una clasificación correcta, a partir de las distancias del punto a clasificar a los puntos base, como se presentó en el ejemplo 1. De esta manera, antes de obtener la etiqueta para un punto a clasificar, se puede proporcionar una probabilidad sobre la certeza de esa etiqueta.

Así, para un determinado conjunto de datos, podemos definir los conceptos de soporte y confianza en una clasificación. Soporte es el porcentaje de puntos sobre los que se puede asegurar con cierta probabilidad una clasificación correcta, y confianza es esa probabilidad. Por ejemplo, para la BD audiology un clasificador estándar proporciona la información de que su tasa de acierto es del 61.9% (k=5 en la tabla 1). Sin embargo, mediante la clase continua podemos establecer un umbral que nos permita variar los valores de soporte y confianza. Así podría ser interesante mejorar esa tasa de error, estableciendo que hay un porcentaje de puntos en los que la probabilidad de esa clasificación correcta es mucho mayor. Concretamente de la tabla 2, se puede extraer la información de que el porcentaje de aciertos puede subir al 76% de los puntos, pero que sólo se puede afirmar ese probabilidad sobre el 62% de los datos. O subir esa probabilidad de éxito al 81% pero sólo podría afirmarse sobre el 57% de los puntos. De esta manera, si un punto a clasificar mantiene una distancia-umbral con la siguiente etiqueta mayor de 0.2 se podría clasificar correctamente en el 81% de los casos.

La tabla 2 presenta para k=5 y mediante 10-validación cruzada, los valores de soporte y confianza que se obtienen al variar el umbral. Evidentemente al aumentar el umbral disminuye el soporte pero aumenta la confianza. Se puede destacar como hay bases de datos donde podemos subir la confianza (teniendo en cuenta los valores de partida de la tabla 1) en la predicción con un soporte no demasiado bajo, por ejemplo sonar, audiology o pima.

BD	para k = 5									
	umbral = 0.02		umbral = 0.05		umbral = 0.1		umbral = 0.15		umbral = 0.2	
	sop.	conf.	sop.	conf.	sop.	conf.	sop.	conf.	sop.	conf.
audiology	90	63	86	66	76	69	62	76	57	81
autos	87	72	85	70	78	71	57	72	54	72
balance-sc	92	93	88	96	72	99	68	100	68	100
cleveland-	96	83	87	83	74	84	61	87	48	89
german_cr	99	74	95	74	71	80	64	81	58	82
glass	91	69	91	69	70	75	68	74	58	78
lymphogra	97	86	97	86	95	87	79	86	74	87
pima_diab	74	78	43	85	2	86	1	100	1	100
primary-tu	82	50	72	51	58	55	44	65	37	72
sonar	97	83	80	89	62	93	34	95	14	96
soybean	98	91	97	92	94	94	90	95	87	96
vehicle	90	75	90	75	64	82	55	86	55	86
vote	99	93	98	93	97	93	95	94	92	96
vowel	98	94	98	94	88	96	81	98	78	98
zoo	99	94	98	93	97	94	88	97	88	97

Tabla 2. Valores de soporte y confianza obtenidos al variar el umbral

5. Conclusiones y trabajos futuros

Con idea de añadir riqueza en problemas de aprendizaje supervisado con etiqueta discreta hemos diseñado un modelo en $\Re^{n-1}$ al que trasladamos las etiquetas de las instancias del conjunto de entrenamiento incluyendo así un preprocesamiento que permite incorporar información que perfeccione la clasificación posterior. Al realizar las pruebas comprobamos que no hay perdida de información en la nueva etiqueta, incluso hay alguna ganancia aunque no es muy significativa. La semántica que proporciona una clase continua puede ser aprovechada para clasificadores donde sea interesante proporcionar una seguridad en la clasificación. El interés por este tratamiento de la información nos hace confiar que un perfeccionamiento en el cálculo de los puntos que extienden la semántica de las etiquetas, conservando la filosofía del modelo, puede significar mejoras sustanciales en otros clasificadores.

Así, para futuros trabajos nos proponemos estudiar la aplicación de distintos clasificadores a los nuevos datos obtenidos de la reetiquetación mediante el método aquí descrito. La información añadida que conlleva la nueva etiqueta y las mayores posibilidades que surgen de trabajar con datos continuos, nos lleva a pensar que modelos de aprendizaje con mejores prestaciones en el caso de clase continua, tales como regresión o redes neuronales, obtendrán mejores resultados de clasificación a partir de los datos modificados que con los originales.

6. Referencias

[1] Blake C., Merz EK. UCI repository of machine learning databases, 1998

[2] Dasarathy BV. Nearest neighbour (NN) Norms: NN pattern classifcation techniques. IEEE Computer Society Press, 1991.

[3] Witten IH and Frank E. "WEKA, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementatios", Morgan Kauffman Publisher. 1999. http://www.cs.waikato.ac.nz/~ml/weka