Una propuesta basada en la estimación de las probabilidades para la edición utilizando el clasificador k-NN

F. Vázquez [1], F. Pla [2], J. S. Sánchez [2]

1. Departamento de Computación
Universidad Oriente
Avda. Patricio Lumumba S/N. 90500 Santiago de Cuba, Cuba
fvazquez@csd.uo.edu.cu

2. Departamento de Lenguajes y Sistemas Informáticos
Universitat Jaume I
Avda. Sos Baynat S/N. 12071 Castellón, España
{sanchez, pla}@lsi.uji.es

Resumen

Los clasificadores supervisados basan su aprendizaje en un conjunto de datos denominado conjunto de entrenamiento, mediante el cual, se proporciona al clasificador una serie de casos o situaciones con las que puede encontrarse al requerirse una predicción o clasificación de un nuevo objeto. La idea central del presente artículo es utilizar como regla de clasificación aquella que utiliza la estimación de la probabilidad de pertenencia a la clase de cada uno de los k-vecinos más cercanos. A partir de este esquema de clasificación se implementa la variante repetitiva del algoritmo de Edición de Wilson utilizando la regla de clasificación de Centroide más Próximo (Wilsoncn), así como también, la de los algoritmos de Wilson con probabilidades y Wilson con probabilidades y umbral. Todos los algoritmos propuestos se comparan con algunas de las más populares técnicas de edición reportadas en la literatura; como son Edición por Partición, Multiedit y algoritmo de Wilson.

Palabras clave: Regla k-NN, Regla k-NCN, Selección de prototipos, Probabilidad por clase.

ingenieriles, computacionales y/o matemáticos relacionados con objetos físicos y/o abstractos, que tienen el propósito de extraer la información que le permita establecer propiedades y/o vínculos o entre conjuntos de dichos objetos sobre la base de los cuales se realiza una tarea de identificación o clasificación.

Dentro del Reconocimiento de Patrones se puede hablar de dos aproximaciones generales, en función del tipo de espacio de representación utilizado y de cómo se estructura la información correspondiente a cada representación. Una de ellas es el Reconocimiento Estadístico de Formas en el cual se asume que el espacio de representación de los objetos posee una estructura de espacio vectorial. Dentro del enfoque Estadístico del Reconocimiento de Formas se suele hacer distinción entre las aproximaciones paramétricas y las no paramétricas. En el primer caso, se asume un conocimiento a priori sobre la forma funcional de las distribuciones de probabilidad de cada clase sobre el espacio de representación de los objetos, por el contrario, las aproximaciones no paramétricas no suponen ninguna forma de las distribuciones de probabilidad sobre el espacio de representación.

Entre los clasificadores estadísticos no paramétricos, podemos señalar las aproximaciones basadas en criterios de vecindad, sobre las cuales se centra básicamente el presente trabajo, que entre sus ventajas respecto a otros métodos no paramétricos, la más inmediata hace referencia a su simplicidad conceptual. La regla del vecino más cercano [1], es uno de los algoritmos más ampliamente estudiados en toda la literatura dentro de los clasificadores estadísticos no paramétricos.

Dado un conjunto de N prototipos previamente etiquetados (conjunto de entrenamiento, TS), esta regla asigna a un objeto que se quiere etiquetar la clase a la cual pertenece su vecino más cercano en el conjunto de entrenamiento de acuerdo a una medida de similaridad en el espacio de rasgos. Otro algoritmo ampliamente estudiado es la regla de los k-vecinos más cercanos en la cual los k-vecinos más cercanos al objeto a clasificar brindan un voto y el objeto es asignado a la clase más votada por sus k-vecinos. En el caso particular que k =1 esta regla coincide con la regla del vecino más cercano.

Un inconveniente que tiene la regla k-NN, es que el conjunto de entrenamiento deberá ser relativamente grande con el fin de garantizar la convergencia del error de clasificación asociado a la regla k- NN, pudiéndose pensar en algunos casos, desechar su utilización, por su elevada complejidad computacional.

En este contexto bajo el nombre genérico de Selección de Prototipos para la regla k-NN se hace referencia a todo un conjunto de procesos previos a la aplicación de los esquemas de clasificación, entre los que podemos señalar, los métodos de Edición y Condensado, dos grupos de técnicas complementarias y fuertemente

relacionadas que, de forma genérica, intentan obtener un conjunto reducido de prototipos sin que existan solapamientos entre regiones de clases distintas. La finalidad de estos procedimientos es aumentar la tasa de aciertos en el proceso de clasificación, así como simplificar la correspondiente carga computacional asociada a la regla de decisión NN.

La aplicación de los procedimientos de edición resultará interesante no sólo como herramienta para disminuir el error de clasificación asociado a las distintas reglas NN, sino también para llevar a cabo cualquier proceso posterior que pudiese beneficiarse de un conjunto de entrenamiento con unas fronteras de decisión sencillas. La idea común a todos los esquemas de Edición consistirá en descartar prototipos que se encuentren en la región correspondiente a alguna clase distinta a la suya, es decir, prototipos cuya probabilidad de pertenencia a su clase se vea superada por la probabilidad de pertenencia a alguna otra clase diferente a la suya [2], [3]. El algoritmo de Edición de Wilson [4] constituye la primera propuesta formal de algoritmo de edición con el objetivo de reducir el conjunto de entrenamiento para la regla NN mediante la eliminación de prototipos erróneamente etiquetados.

La idea central del presente artículo es utilizar como regla de clasificación aquella que utiliza la estimación de la probabilidad de pertenencia a la clase de cada uno de los k-vecinos más cercanos. A partir de este esquema de clasificación se implementa la variante repetitiva del algoritmo de Edición de Wilson que utiliza como regla de clasificación la de Centroíde más Próximo (Wilsoncn) [5], así como también, la de los algoritmos de Wilson con probabilidades y Wilson con probabilidades y umbral, [6]. Los algoritmos propuestos se comparan con algunas de las más populares técnicas de edición reportadas en la literatura; como son Edición por Partición [7], Multiedit [7] y algoritmo de Wilson [4].

Para comparar estos algoritmos usamos bases de datos reales tomadas del UCI Machine Learning Database Repository [8].

2 Algoritmos de Edición

Todos los algoritmos que mostraremos en este trabajo se basan en el clasificador k-NN. Esta regla puede expresarse formalmente de la siguiente manera.

Sea $\{X, \theta\} = \{ (x_1, \theta_1), (x_2, \theta_2) , \ldots , (x_n, \theta_n)\}$ un conjunto de entrenamiento con N prototipos de M posibles clases, y sea $P_j = \{P_{j,i} / i = 1,2,..., N_j \}$ el conjunto de prototipos de X pertenecientes a la clase j. Se puede definir la Vecindad $V_k(x)$ de una muestra $x \in X$ como el conjunto de prototipos que cumple con las condiciones:

$$V_k(x) \subseteq P \; ; \; |V_k(x| = k$$

$$\forall \; p \in V_k(x), q \in P\text{-}\, V_k(x) \Rightarrow d(p,x) \leq d(q,x) \; ; \quad \text{donde} \quad P = \bigcup_{i=1}^{M} P_i$$

Si ahora definimos una nueva distancia entre un punto y un conjunto de prototipos tal como:

$$d_k(x,P_i) = k - \left| V_k(x) \cap P_i \right|$$

Podremos definir la regla de clasificación k-NN de la siguiente manera:

$$\delta_{k\text{-}NN}(x) = \Theta_i \Leftrightarrow d(x,P_i) = \min_{i=1,2,\ldots,M} d_k(x,P_j)$$

El significado de esta expresión se puede resumir en que la clase asignada a la muestra x será la clase más votada entre los k vecinos más próximos del conjunto de entrenamiento.

2.1 Algoritmo de Edición de Wilson

Como habíamos expresado anteriormente la idea común de todos los algoritmos de Edición consistirá en descartar prototipos que se encuentren en la región correspondiente a una clase distinta a la suya, es decir, prototipos cuya probabilidad de pertenencia a su clase se vea superada por la probabilidad de pertenencia a alguna otra clase.

El algoritmo de edición de Wilson [4] constituye la primera propuesta formal con el objetivo de reducir el conjunto de entrenamiento para la regla NN mediante la eliminación de prototipos erróneamente etiquetados. La idea de este método radica en que si un prototipo es erróneamente clasificado usando la regla k-NN es eliminado del conjunto de entrenamiento (TS), para este fin, se utilizarán todos los prototipos del conjunto de entrenamiento para determinar los k-vecinos más próximos (excepto el prototipo que se esta considerando en cada momento), es decir, el método de estimación del error empleado corresponderá al leaving-one-out. El algoritmo de edición de Wilson se expresa de la siguiente manera.

1. Inicialización $S \leftarrow X$
2. Para cada prototipo $x_i \in X$
   2.1. Buscar los k vecinos más próximos de x_i en $X-\{x_i\}$
   2.2. Si $\delta_{k\text{-}NN}(x_i) \neq \theta_i$, entonces $S \leftarrow S-\{x_i\}$

El algoritmo de Edición de Wilson proporciona un conjunto de prototipos organizados en agrupamientos relativamente compactos y homogéneos, es preciso señalar que el coste computacional de este algoritmo de edición es de $O(N^2)$, lo cual puede ocasionar ciertos problemas prácticos para conjuntos de entrenamiento relativamente grandes.

2.2 Algoritmo de Edición por partición (X,k,m)

En este esquema de edición, el método de estimación de la pertenencia de un prototipo a su clase consistirá, en realizar una partición del conjunto de entrenamiento en m bloques disjuntos de prototipos y, después de enumerarlos, hacer una estimación para cada bloque j utilizando el bloque ((j+1) módulo m) para diseñar el clasificador. Este procedimiento se puede considerar estadísticamente independiente siempre que m>2, el mismo recibe el nombre en la literatura de Algoritmo de Edición Holdout [7]

Un resumen de esté algoritmo se muestra a continuación

1. Hacer una partición aleatoria de X en m bloques $T_1,...,T_m$
2. Para cada bloque T_j (j=1,...,m):
   2.1. Para cada $x_i \in T_j$
      2.1.1. Buscar los k vecinos más próximos de x_i en $T_{((j+1) \bmod m)}$
      2.1.2. Si $\delta_{k\text{-}NN}(x_i) \neq \theta_i$ hacer $X \leftarrow X-\{x_i\}$

2.3 Algoritmo Multiedit

El algoritmo Multiedit [7], consiste en repetir la edición por partición pero utilizando la regla NN con (k=1). Este algoritmo se puede presentar de manera resumida de la siguiente forma:

1. Inicialización: $t \leftarrow 0$
2. Repetir hasta que en las últimas t iteraciones (t>f) no se produzcan ninguna eliminación de prototipos del conjunto X.
   2.1. Asignar a S el resultado de aplicar la Edición por Partición sobre X utilizando la regla NN
   2.2. Si no se ha producido ninguna nueva eliminación en el paso 2.1 ($|X| = |S|$), hacer $t \leftarrow t+1$ e ir al paso 2.
   2.3. Asignar a X el contenido de S y hacer $t \leftarrow 0$

Debemos de señalar que la gran ventaja de la versión iterativa radica en que para conjuntos finitos aunque suficientemente grandes, su comportamiento resulta ser

significativamente mejor por el hecho de no presentar aquella dependencia del parámetro k que mostraba el anterior algoritmo.

2.4 Regla de los k-vecinos de Centroide más cercanos

En [5] se propone una nueva definición de vecindad, la cual tiene en cuenta los criterios de distancia y simetría, la misma es aplicada sobre problemas generales de clasificación, esta regla recibe el nombre de vecindad de centroide más próximo (Nearest Centroid Neighborhood, NCN). La formulación de este concepto es como sigue:

Sea $X = \{x_1, x_2, ..., xn\}$ un conjunto de objetos, y sea p un cierto punto al que queremos encontrar sus k-vecinos de centroide más próximos, con este fin seguiremos el siguiente procedimiento iterativo, en el que el primer vecino del punto p corresponde a su vecino más próximo, mientras que los sucesivos vecinos se tomarán de manera que minimicen la distancia entre p y el centroide de todos los vecinos seleccionados hasta el momento. Así, si calculamos el k-ésimo vecino a partir de los k-1 vecinos previamente elegidos por el principio de centroide más próximo conseguiremos cumplir con los criterios de distancia y simetría.

Debemos señalar que, como consecuencia del criterio de centroide que se está utilizando, todos los k-NCN vecinos seleccionados se situarán alrededor del punto p, es decir, de alguna forma que se consigue que dicho punto quede envuelto por sus k-vecinos.

Valiéndose de esta definición de vecindad en [5] se propone la siguiente regla de clasificación que se denominó regla de los k-Vecinos de Centroides más Cercanos (k-NCN), la cual se puede formalizar de la siguiente manera.

$$\delta_{k\text{-}NCN}(x) = \varpi_i \Leftrightarrow d(x, P_i) = \min_{i=1,2,...,M} d_k(x, P_j)$$

Esta expresión significa, que la clase asignada a la muestra x corresponderá a la clase más votada entre los k vecinos de centroide más próximo. En la práctica, al igual que ocurría con la regla de decisión k-NN, deberíamos considerar siempre un número impar de vecinos con el fin de evitar posibles empates.

Aplicando esta regla de clasificación surge una variante del algoritmo de Wilson que se denomina en la Literatura Wilsoncn el cual se expresa de la siguiente manera.

1. Inicialización S ←X
2. Para cada prototipo $x_i \in X$
   2.1. Buscar los k vecinos de centroide más próximos de x_i en X-$\{x_i\}$
   2.2. Si $\delta_{k\text{-}NCN}(x_i) \neq \theta_i$, entonces S ← S-$\{x_i\}$.

3 Algoritmos de Edición utilizando probabilidades de clases

Todos los algoritmos de edición descritos en la sección anterior utilizan como regla de edición la regla K-NN o la regla k-NCN, en [6] se propone una nueva regla de edición la cual tiene en cuenta la probabilidad de un objeto de pertenecer a una cierta clase.

Para este fin primeramente se definió la función estrictamente positiva.

$$P_i(x) = \sum_{j=1}^{k} p_i^j \frac{1}{(1 + d(x, x^j))}$$

donde p_i^j denota la probabilidad de pertenencia del j-ésimo vecino más cercano x^j de pertenecer a la clase i. Es necesario señalar que los objetos que pertenecen a la clase i su probabilidad de pertenencia a esta clase será 1 y la probabilidad de pertenecer a una clase distinta a la suya será cero.

Posteriormente para calcular la probabilidad de un objeto de pertenecer a una determinada clase se calcula valiéndonos de la siguiente expresión:

$$p_i(x) = P_i(x) / \sum_{j=1}^{M} P_j(x)$$

Teniendo en cuenta las dos expresiones expuestas anteriormente se define la siguiente regla de clasificación:

$$\delta_{k\text{-}prob}(x) = i \; ; \quad i \,/\, p_i(x) = \arg\max_j (p_j(x))$$

Donde la expresión significa, que el objeto x es asignado a la clase de mayor probabilidad. Con esta nueva regla de edición dos algoritmos de edición fueron descritos en [6], WilsonProb y WilsonUmb que describimos a continuación:

Algoritmo de Edición WilsonProb

1. Inicialización: $S \leftarrow X$
2. Para cada prototipo $x \in X$
   2.1. Buscar los k vecinos más próximos en X - {x}
   2.2. Si $\delta_{k\text{-}prob}(x) \neq \theta$, hacer $S \leftarrow S-\{x\}$, donde θ denota la clase a la cual pertenece el objeto x.

Algoritmo de Edición WilsonUmb

1. Inicialización: $S \leftarrow X$
2. Para cada prototipo $x \in X$
   2.1. Buscar los k vecinos más próximos en X - {x}
   2.2. Si $\delta_{k\text{-}prob}(x) \neq \theta$ ó $p_i \leq \mu$, hacer $S \leftarrow S - \{x\}$, θ denota la clase a la cual pertenece el objeto x y p_i es la mayor de todas las probabilidades de clase.

Siguiendo este mismo esquema, es decir, utilizando como regla de clasificación la regla de los k vecinos de centroide más cercano (k-NCN) y además teniendo en cuenta la probabilidad de pertenencia a la clase de cada uno de los vecinos, se hace una modificación del algoritmo de Edición de Wilsoncn, el mismo será denotado como Wilsoncn-prob, el cual, describimos de la siguiente manera.

1. Inicialización $S \leftarrow X$
2. Para cada prototipo $x_i \in X$
   2.1. Buscar los k vecinos más próximos de x_i en $X-\{x_i\}$
   2.2. Si $\delta_{k\text{-}NCN\text{-}prob}(x_i) \neq \theta_i$, entonces $S \leftarrow S-\{x_i\}$.

4 Edición repetitiva utilizando probabilidades de clases

Puestos que los algoritmos de edición en sentido general proporcionan un conjunto de prototipos organizados en grupos más o menos compactos y homogéneos, cabría esperar que la repetición de este procedimiento fuese capaz de potenciar aún más dicho efecto de hecho este argumento fue propuesto por [9], para proponer el siguiente algoritmo de edición.

1. Inicialización $S \leftarrow \varnothing$
2. Mientras $|X| \neq |S|$
   2.1. Pasar el contenido actual de X a S: $S \leftarrow X$
   2.2. Aplicar Edición de Wilson sobre el conjunto X

Esta es la idea fundamental que hemos seguido en este trabajo, es decir, aplicar de forma repetitiva todos los algoritmos de Edición cuya regla de edición sea la que no solo tiene en cuenta a los vecinos más cercanos, o a los vecinos de centroide más cercanos sino que además tenga en cuenta la probabilidad de pertenencia a la clase de estos vecinos. De manera general estos algoritmos se pudieran escribir de la siguiente manera.

1-Inicialización S ←∅
2-Mientras | X| ≠| S|
 2.1-Pasar el contenido actual de X a S: S←X
 2.2-Aplicar Edición con probabilidades de clase sobre el conjunto X.

En este trabajo hemos implementado la variante repetitiva del algoritmo Wilsoncn-prob, la cual se puede resumir mediante los siguientes pasos:

1-Inicialización S ←∅
2-Mientras | X| ≠| S|
 2.1-Pasar el contenido actual de X a S: S←X
 2.2- Aplicar Edición de Wilsoncn-prob sobre el conjunto X .

Similarmente también se muestran los algoritmos repetitivos de Wilson-prob y Wilson-prob-umb, los que de manera resumida se pueden mostrar como siguen:

1-Inicialización S ←∅
2-Mientras | X| ≠| S|
 2.1-Pasar el contenido actual de X a S: S←X
 2.2- Aplicar Edición de Wilson-prob (Wilson-prob –umb) sobre el conjunto X .

5 Resultados experimentales y discusión

Para llevar a cabo los experimentos se utilizaron conjuntos o bases de datos obtenidos del repositorio Machine Learning Database de la Universidad de California, Irving [8]. De las bases de datos disponibles en el repositorio, para esta fase experimental, se eligieron 10 bases de datos reales. Las principales características de estas bases de datos se muestran en la Tabla 1.

En los experimentos realizados con cada una de las bases de datos, utilizamos el método de validación cruzada, dividiendo el conjunto original en 5 conjuntos de entrenamiento y 5 conjuntos de prueba, el (80% de los objetos del conjunto original fueron tomados para el conjunto de entrenamiento TS y el 20% de los restantes para el conjunto de prueba), los mismos fueron empleados para estimar

los porcentajes de clasificación, así como también para estimar la reducción de la talla de los conjuntos de entrenamiento. En cada una de las tablas que aparecen en este trabajo, mostramos para cada una de las bases de datos dos filas, la primera corresponde a los porcentajes de clasificación y la segunda a la reducción del conjunto de entrenamiento, en cada tabla se señala utilizando letras negritas o letras cursivas las características que queremos resaltar de cada uno de los algoritmos presentados.

	No. clases	No. rasgos	No. objetos
Cancer	2	9	683
Liver	2	6	345
Heart	2	13	270
Wine	3	13	178
Australian	2	42	690
German	2	24	1002
Ionosphere	2	34	352
Phoneme	2	5	5404
Satimage	6	36	6453
Textura	11	40	5500

Tabla 1. Características de las bases de datos

	NN	Wils.	Hold.	Mult.	Wilson	Wil-pro	Wil-pro-rep
					k-NCN	k-NCN	k-NCN
Cancer	95.60	96.19	**96.63**	**96.63**	95.89	96.04	96.33
		3.44	4.28	*7.43*	3.00	3.00	3.95
Liver	65.79	70.70	70.40	59.49	**71.03**	70.02	70.92
		32.89	37.10	*75.79*	32.46	34.56	50.07
Heart	58.16	67.00	67.34	66.64	69.21	**69.95**	68.47
		34.44	38.70	*69.25*	31.94	38.61	44.53
Wine	73.04	70.90	**75.24**	72.42	72.55	72.07	72.75
		34.97	30.75	*45.50*	25.70	22.89	34.41
Ionosphere	83.46	82.02	82.31	69.58	**84.55**	84.14	84.41
		16.66	14.52	34.11	23.08	27.76	*37.67*
Australian	65.67	69.27	**70.72**	68.99	68.26	68.83	68.83
		31.88	36.88	*59.52*	30.03	31.26	*42.42*
German	64.81	70.40	72.00	70.00	**72.6**	71.58	71.02
		30.50	32.27	*54.72*	26.9	27.22	36.67

Tabla 2. % de clasificación y reducción de la talla del conjunto de entrenamiento

En los resultados que aparecen en la Tabla 2, se muestran los diferentes métodos de edición anteriormente descritos en el trabajo, donde hacemos hincapié en los resultados obtenidos por los algoritmos Wilsoncn-prob y Wilsoncn-prob-rep,

donde se puede apreciar que los mismos obtienen similares resultados a los obtenidos por los demás algoritmos y los porcentajes de reducción de la talla del conjunto de entrenamiento de estos algoritmos en todos los casos supera al algoritmo Wilsoncn, con letras negritas se muestran los mejores algoritmos en cuanto a porcentajes de clasificación, así como también, mostramos en letras cursivas, los mejores porcentajes de reducción de la talla del conjunto de entrenamiento.

	NN	Wils.	Hold.	Mult.	WProb	Wil-prob-rep
Cancer	95.60	96.19	96.63	96.63	96.34	*96.48*
		3.44	4.28	7.43	3.36	*4.24*
Liver	65.79	70.70	70.40	59.49	68.67	*68.66*
		32.89	37.10	75.79	27.89	*45.00*
Heart	58.16	67.00	67.34	66.64	66.26	*65.92*
		34.44	38.70	69.25	28.51	*43.79*
Wine	73.04	70.90	75.24	72.42	69.69	*68.61*
		34.97	30.75	45.50	14.60	*33.98*
Ionosphere	83.46	82.02	82.31	69.58	81.74	*81.15*
		16.66	14.52	34.11	18.01	*33.04*
Texture	98.96	98.63	98.56	94.62	98.74	*98.74*
		1.34	3.69	15.31	1.01	*2.65*
Australian	65.67	69.27	70.72	68.99	69.56	*68.55*
		31.88	36.88	59.52	25.90	*41.41*
German	64.81	70.40	72.00	70.00	70.70	*70.70*
		30.50	32.27	54.72	26.90	*26.9*
Phoneme	70.26	73.53	74.29	75.35	73.42	*73.70*
		10.56	16.07	37.43	11.98	*15.59*
Satimage	83.62	83.29	83.32	82.35	83.09	*82.91*
		9.43	10.19	24.51	9.25	*12.86*
Glass	71.4	70.40	66.03	58.63	70.70	*70.30*
		30.5	46.14	61.21	26.9	*41.72*
Balance	79.20	85.11	85.62	86.41	84.96	*84.80*
		14.8	14.52	37.04	10.76	*14.84*

Tabla 3. % de clasificación y reducción de la talla del conjunto de entrenamiento

En la Tabla 3, mostramos los resultados experimentales obtenidos por el algoritmos Wilson-prob-rep sobre las 10 base de datos, estos resultados han sido obtenido sobre las 5 particiones en las cuales fueron divididas cada una de las base de datos, el la tabla anteriormente descrita se muestran con letras cursivas los valores correspondientes a porcentajes de clasificación, así como también los porcentajes de reducción de la talla del conjunto de entrenamiento, donde podemos señalar que aunque los porcentajes de clasificación obtenidos por este algoritmo, no alcanza superar los resultados que obtuvo el algoritmo Wilson-prob, los porcentajes de

clasificación del mismo son bastantes similares a todos los algoritmos, superando Wilson-prob-rep siempre los porcentajes en la reducción del conjunto de entrenamiento.

	NN	Wils.	Wil-pro	Wilson-prob-umb			Wilson-prob-umb		
				0.6	0.6-r	0.7	0.7-r	0.8	0.8-r
Cancer	95.60	96.19	96.34	96.48	*96.48*	96.63	***96.78***	96.78	***96.78***
		3.44	3.36	4.09	*5.45*	5.49	*8.38*	*7.68*	*8.63*
Liver	65.79	70.70	68.67	68.97	*69.82*	*69.55*	*69.54*	*68.95*	*65.18*
		32.89	27.89	45.94	*63.11*	61.37	*71.37*	*67.82*	*71.73*
Heart	58.16	67.00	66.26	65.17	*65.94*	65.12	*66.32*	64.78	*65.63*
		34.44	28.51	40.09	*57.40*	53.61	*68.14*	65.00	*74.21*
Wine	73.04	70.90	69.69	69.74	*70.12*	69.20	*71.84*	*69.20*	*67.44*
		34.97	14.60	33.28	*44.24*	35.67	*47.47*	41.43	*48.23*
Ionosp	83.46	82.02	81.74	*81.74*	*81.15*	*80.89*	*80.28*	*81.16*	*81.15*
		16.66	18.01	18.01	*17.23*	24.21	*35.47*	25.21	*37.23*
Textur	98.96	98.63	98.74	*98.49*	*98.20*	*98.29*	*98.02*	*98.32*	*98.02*
		1.34	1.01	1.50	*2.77*	3.17	*4.31*	3.06	*4.32*
Austra	65.67	69.27	69.56	*69.70*	*68.84*	68.39	*68.85*	68.54	*68.55*
		31.88	25.90	37.02	*49.02*	50.76	*61.88*	57.53	*62.53*
German	64.81	70.40	70.70	*71.10*	***70.50***	*70.50*	*70.20*	*70.50*	*70.30*
		30.50	26.90	39.62	*45.5*	*52.72*	*65.02*	*60.00*	*67.9*
Phonem	70.26	73.53	73.42	73.44	*73.97*	74.02	*74.34*	73.99	***74.42***
		10.56	11.98	17.26	*16.86*	24.36	*24.90*	29.15	*25.31*
Satima	83.62	83.29	83.09	*83.18*	*83.04*	*84.24*	*82.91*	*83.50*	*82.90*
		9.43	9.25	15.61	*17.48*	19.22	*18.41*	23.90	*21.80*

Tabla 4. % de clasificación y reducción de la talla del conjunto de entrenamiento

En la Tabla 4 se muestra específicamente una comparación entre los algoritmos de Wilson, Wilson-prob y Wilson-prob-rep tomando el umbral los valores 0.6, 0.7 y 0.8, los valores tantos de los porcentajes de clasificación como los porcentajes de reducción para los diferentes valores del umbral en el caso del algoritmo Wilson-umb-rep lo hemos indicado con la letras cursivas, y hemos señalado en negrita los mejores porcentajes de clasificación de cada algoritmo en cada una de las bases de datos, es preciso volver a señalar que las variantes repetitivas para los diferente valores del umbral siempre obtienen mejores porcentajes de reducción de la talla del conjunto de entrenamiento

6 Conclusiones y trabajo futuro

En este trabajo especialmente hemos utilizado una regla de edición que no solo tiene en cuenta la cercanía de los vecinos, sino además de esto se ha tenido en

cuenta la probabilidad de cada uno de los vecinos de pertenecer a una clase dada. Con esta idea, mostramos una modificación del algoritmo de Wilsoncn, así como también en todos los casos hemos implementado las variantes repetitivas de todos los algoritmos que tienen como regla de edición la de probabilidades de clases. Una serie de experimentos con bases de datos reales muestras los resultados obtenidos. En los mismos se puede comprobar que los resultados alcanzados por estos algoritmos mantienen similares porcentajes de clasificación que los tradicionales algoritmos que se presentan en la literatura, pero superan a los mismos en cuanto a porcentajes de reducción del conjunto de entrenamiento

Estos métodos de edición en los que como regla de edición se utiliza la probabilidad de pertenencia a la clase de sus vecinos serán utilizados en futuros trabajos de aprendizaje parcialmente supervisados, donde para clasificar a un nuevo objeto no solo se use un conjunto de entrenamiento, sino que este a su vez vaya enriqueciendo su conocimiento mediante objetos que han sido etiquetados por el mismo algoritmo.

Referencias

[1] Dasarathy, B. V.: Nearest Neighbor Norms: NN Pattern Classification Techniques. IEEE Computer Society Press, Los Alamos, CA (1991)

[2] Sánchez, J.S., Barandela, R., Marqués, A.I., Alejo, R., Badenas, J. : Analysis of new techniques to obtain quality training sets. Pattern Recognition Letters 24 (2003) 1015-1022.

[3] Wilson, D.R., Martinez, T.R.: Reduction techniques for instance-based learning algorithms. Machine Learning 38 (2000) 257-286.

[4] Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data sets. IEEE Trans. on Systems, Man and Cybernetics 2 (1972) 408-421.

[5] Sánchez,J.S., F,Pla and F.J.Ferri, Using the nearest centroid neighbourhood concept for editing purpose,In Proc. VII Simposium Nacional de Reconocimiento de formas y Análisis de Imágenes 1, 175-180 (1997).

[6] Vázquez,F.;Sánchez, J.S.; Pla, F.;"A stochastic approach to Wilson's editing algorithm", Pattern Recognition and Image Analysis, Lecture Notes in Computer Science, Vol. 3523, J.S. Marques, N. Pérez de la Blanca and P. Pina (Eds.), Springer-Verlag, ISBN 3-540-26154-0, pp. 35-42, 2005.

[7] Devijver, P.A., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice Hall, Englewood Cliffs, NJ (1982).

[8] Merz, C.J., Murphy., Murphy, P.M.: UCI Repository of Machine Learning Database. Department of Information and Computer Science, University of California, Irvine, CA (1998).

[9] Tomek,I., An experimental with the edited nearest neighbor rule, IEEE Tans. on Systems, Man and Cybernetics SMC-6, 448-452 (1976).