

La Teoría de los Conjuntos Aproximados y las Técnicas de Bootstrap para la Edición de Conjuntos de Entrenamiento. Su Aplicación en el Pronóstico Meteorológico

Rough Set Theory and Bootstrap's Technique to Edit Training Set. The Application in Seasonal Weather Forecasting

Beitmantt Cárdenas, MSc.¹, Yailé Caballero, PhD.², Rafael Bello, PhD.²

¹Grupo de Investigación GTecni-IA, Universidad Pedagógica y Tecnológica de Colombia

²Departamento de Computación, Universidad de Camagüey, Cuba
{beitmantt, yailec}@yahoo.com, rbellop@uclv.edu.cu

Recibido para revisión 20 de Febrero de 2007, Aceptado 03 de Diciembre, Versión final 14 de Diciembre de 2007

Resumen— Un conjunto de entrenamiento es una muestra de datos para el entrenamiento de clasificadores. Estos datos, están expuestos al adquirir su conocimiento, a tener prototipos erróneamente etiquetados, lo que implica un aprendizaje infructuoso en su aplicación. La selección de los objetos de un dominio a incluir en un conjunto de entrenamiento es un problema presente en todos los modelos computacionales que realizan inferencias a partir de ejemplos. La edición de un conjunto de entrenamiento se hace con el objetivo de eliminar los prototipos que inducen a una incorrecta clasificación supervisada, seleccionando un conjunto de referencia representativo y reducido. Las técnicas de edición, también producen la eliminación de prototipos, y con ello, la reducción de la matriz de aprendizaje. En este trabajo se propone un nuevo algoritmo con estos propósitos EDITBRS. Se validan los resultados del mismo en el problema real del pronóstico de las temperaturas en la estación meteorológica de la U.P.T.C., Colombia.

Palabras Clave— Edición de Conjuntos de Entrenamiento, Técnicas de Bootstrap, Teoría de los Conjuntos Aproximados, Pronóstico de Variables Meteorológicas.

Abstract—Rough Set Theory (RST) is a technique for data analysis. In this study, we use RST and bootstrap's technique to improve the performance of classifiers. The RST is used to edit and reduce the training set. We propose a method to edit training sets, which is based on the lower and upper approximations and bootstrap's technique. The accelerated growth of the environmental of information volumes on processes, phenomena and reports brings about an increasing interest in the possibility of discovering knowledge from data sets. Experimental results show a satisfactory performance using these techniques.

Keywords— Bootstrap's Technique, Edit to Training Set, Rough Set Theory, Weather forecasting.

I. INTRODUCCIÓN

UN problema que afecta seriamente a los métodos de clasificación supervisada es la presencia de prototipos erróneamente etiquetados en el conjunto de entrenamiento. Durante el entrenamiento pueden producirse errores en el etiquetado o pueden aparecer patrones ruidosos por problemas en la recopilación de los datos. Estos prototipos suelen aparecer en zonas cercanas a las regiones de decisión e influyen negativamente en el entrenamiento ya que pueden disminuir la precisión de la clasificación supervisada. Además, existe un alto costo computacional asociado a la aplicación de los métodos de clasificación supervisada al conjunto completo de prototipos. Según [1] los métodos de edición se pudieran agrupar en i) Métodos de selección de objetos (aquellos en los cuales solo se seleccionan objetos del conjunto de entrenamiento siguiendo cierto criterio) y ii) Métodos de construcción de objetos (aquellos en los que los nuevos prototipos pueden no estar en la matriz de entrenamiento).

En la sección 2 de este artículo se mencionan brevemente algoritmos que se han reportado para editar conjuntos de entrenamiento; en la sección 3 se propone un nuevo método de edición de conjuntos de entrenamiento a través de la Teoría de los Conjuntos Aproximados y en la sección 4 se validan los resultados y se describe la aplicación de este método en un problema real.

II. EDICIÓN DE CONJUNTOS DE ENTRENAMIENTO

A. Métodos de selección de objetos

Se han reportado varios métodos de selección de objetos: Condensed Nearest Neighbors (CNN) [2]; Reduced Nearest Neighbor (RNN) [3]; Edited Nearest Neighbor (ENN) [4]; All-kNN [5]; Multiedit [6]. La desventaja de estos métodos consiste en que los dos primeros son dependientes del orden de presentación de los objetos. Según [7] el método Multiedit trabaja bien en grandes conjuntos de datos; sin embargo, para conjuntos de datos pequeños o con clases entremezcladas, puede eliminar todos los objetos de una clase y si las clases están perfectamente separadas, prácticamente no elimina objetos. Aha [8] presentó una serie de algoritmos de aprendizaje basado en instancias (Instance-Based Learning, IBL), estos pueden retener objetos ruidosos. En los últimos cinco años se han reportado varios métodos de edición: ICF [9]; dos esquemas para reducir el tiempo de ejecución del método BSE [10]; CSE [11]. La desventaja principal de este último radica en que no logra niveles de compactación comparables con otros métodos, incluso en clases totalmente separadas [1]. En [12] se propone el algoritmo de edición Edit1RS basado en el concepto de aproximación inferior de la Teoría de los Conjuntos Aproximados, con el que se obtienen resultados satisfactorios.

B. Métodos de construcción de objetos

Koplowitz y Brown en 1978 [13] hicieron una modificación del algoritmo de Wilson, le llamaron Generalized Editing (GE). Ellos estaban preocupados con la posibilidad de remover demasiados prototipos del conjunto de entrenamiento en el procedimiento de la Edición de Wilson. Este método consiste en eliminar algunos de los prototipos y además puede cambiar etiquetas de las clases de otros objetos. Se puede considerar como una técnica para modificar la estructura de la muestra de entrenamiento (a través de re-etiquetar algunos de los objetos de entrenamiento) y no solamente para eliminar objetos atípicos [14]. Hamamoto y colaboradores en 1997 [15] desarrollaron cuatro métodos de Bootstrap [16]. Los objetos son construidos por la media de los k vecinos más cercanos de su misma clase, de np objetos seleccionados al azar (sin repeticiones). El usuario define el número de intentos que se realizarán, de los que selecciona el mejor. Según Bezdek [17] la familia de Bootstraps permite obtener muy buenos resultados.

III. NUEVO MÉTODO DE EDICIÓN DE CONJUNTOS DE ENTRENAMIENTO BASADO EN LA TEORÍA DE LOS CONJUNTOS APROXIMADOS

A. Consideraciones necesarias sobre la Teoría de los Conjuntos Aproximados

La Teoría de Conjuntos Aproximados (Rough Sets Theory) fue introducida por Z. Pawlak en 1982 [18]. Se basa en aproximar cualquier concepto, un subconjunto duro del

dominio como por ejemplo, una clase en un problema de clasificación supervisada, por un par de conjuntos exactos, llamados aproximación inferior y aproximación superior del concepto. Con esta teoría es posible tratar tanto datos cuantitativos como cualitativos, y no se requiere eliminar las inconsistencias previas al análisis. La inconsistencia describe una situación en la cual hay dos o más valores en conflicto para ser asignados a una variable [19, 20].

Los conceptos básicos de la RST son las aproximaciones inferiores y superiores de un subconjunto $X \subseteq U$. Estos conceptos fueron originalmente introducidos con referencia a una relación de inseparabilidad R . Sea R una relación binaria definida sobre U la cual representa la inseparabilidad, se dice que $R(x)$ significa el conjunto de objetos los cuales son inseparables de x . Así, $R(x) = \{y \in U : yRx\}$. En la RST clásica, R es definida como una relación de equivalencia; es decir, es una relación binaria $R \subseteq U \times U$ que es reflexiva, simétrica y transitiva. R induce una partición de U en clases de equivalencia correspondiente a $R(x)$, $x \in U$.

Este enfoque clásico de RST es extendido mediante la aceptación que objetos que no son inseparables pero sí suficientemente cercanos o similares puedan ser agrupados en la misma clase [21]. El objetivo es construir una relación R' a partir de la relación de inseparabilidad R pero flexibilizando las condiciones originales para la inseparabilidad. Esta flexibilización puede ser realizada de múltiples formas, así como pueden ser dadas varias definiciones posibles de similitud. Existen varias funciones de comparación de atributos (funciones de similitud), las cuales están asociadas al tipo del atributo que se compara [22, 23]. Sin embargo, la relación R' debe satisfacer algunos requerimientos mínimos. Si R es una relación de inseparabilidad definida en U , R' es una relación de similitud extendida de R si y solo si $\forall x \in U, R(x) \subseteq R'(x)$ y $\forall x \in U, \forall y \in R'(x), R(y) \subseteq R'(x)$, donde $R'(x)$ es la clase de similitud de x , es decir, $R'(x) = \{y \in U : yR'x\}$. R' es reflexiva, cualquier clase de similitud puede ser vista como un agrupamiento de clases de inseparabilidad y R' induce un cubrimiento de U [24]. Esto muestra que un objeto puede pertenecer a diferentes clases de similitud simultáneamente, lo que significa que el cubrimiento inducido por R' sobre U no es necesariamente una partición.

La aproximación de un conjunto $X \subseteq U$, usando una relación de inseparabilidad R , ha sido inducida como un par de conjuntos llamados aproximaciones R -inferior y R -superior de X . Se considera en esta propuesta una definición de aproximaciones más general, la cual maneja cualquier relación reflexiva R' . Las aproximaciones R' -inferior ($R'_*(X)$) y R' -superior ($R'^*(X)$) de X están definidas respectivamente como se muestra en las expresiones 1 y 2.

$$R_*(X) = \{x \in X : R(x) \subseteq X\} \quad (1)$$

$$R^*(X) = \bigcup_{x \in X} R(x) \quad (2)$$

Teniendo en cuenta las expresiones definidas en 1 y 2, se define la región límite de X para la relación R' [25]:

Si el conjunto BNB es vacío entonces el conjunto X es exacto respecto a la relación R'. En caso contrario, $BN_B(X) \neq \emptyset$, el conjunto X es inexacto o aproximado con respecto a R'.

$$BN_B(X) = R^*(X) - R_*(X) \quad (3)$$

Función de pertenencia aproximada

Esta función cuantifica el grado de solapamiento relativo entre R'(x) (clase de similitud de x) y la clase a la cual el objeto x pertenece. Se define como sigue:

$$\mu_x(x) = \frac{|X \cap R(x)|}{|R(x)|} \quad (4)$$

La función de pertenencia aproximada puede ser interpretada como una estimación basada en frecuencias de $\Pr(x \in X | x, R'(x))$, es decir, la probabilidad condicional de que el objeto x pertenezca al conjunto X [26].

B. EditBRS: un nuevo método de edición sobre la base de la Teoría de los Conjuntos Aproximados

En el método Edit1RS [12] se toman en cuenta sólo los elementos que están en las aproximaciones inferiores. Sin embargo, la información de los objetos que están en la frontera (BNB), puede ser interesante, pero habría que cambiar la clase a aquellos que lo requieran, esto pudiera hacerse a través de un proceso de re-etiquetamiento.

Por su parte, el algoritmo de edición Edit2RS [12] construye objetos a partir de re-etiquetar aquellos que pertenecen al conjunto frontera. Con este método se obtienen resultados satisfactorios; sin embargo, presenta la deficiencia de que el resultado que arroja el algoritmo Generalized Editing depende del orden de los objetos, pues se van analizando las clases de estos según las de sus k-Vecinos más Cercanos.

La función de pertenencia aproximada [26], descrita en la expresión 4, arroja un valor de pertenencia de un objeto a una clase determinada. Si se le aplica esta función a cada uno de los objetos del conjunto frontera para cada una de las clases, se puede saber por cada objeto, a cuál de las clases él tiene mayor grado de pertenencia. Esto constituye una forma de re-etiquetar las clases de los objetos del conjunto frontera. El siguiente algoritmo está propuesto tomando en cuenta estas ideas.

Algoritmo EditBRS:

P1. Construir el conjunto B, $B \subseteq A$. Se sugiere que B sea un reducto del sistema de decisión para disminuir la dimensionalidad de los atributos sin afectar el proceso de clasificación.

P2. Formar los conjuntos $X_i \subseteq U$, tal que todos los elementos del universo (U) que tienen valores d_i en el atributo de decisión están en X_i .

P3. $S_E = \emptyset$.

P4. Para cada conjunto X_i :

Calcular su aproximación inferior ($B_*(X_i)$) y su aproximación superior ($B^*(X_i)$).

p4.1. $S_E = S_E \cup B_*(X_i)$.

p4.2. $T_i = B^*(X_i) - B_*(X_i)$.

P5. Calcular la unión de los conjuntos T_i , para obtener $T = \bigcup T_i$.

P6. Para cada elemento del conjunto T:

p6.1 Calcular la función de pertenencia aproximada (expresión 4) a cada clase. Se utiliza la técnica de Bootstrap[27] para el cálculo de $center(X)$

p6.2 Si la clase con mayor valor (según p6.1) difiere de la clase del objeto, entonces re-etiquetarla por esta clase con la cual se obtuvo mayor valor de pertenencia del objeto.

P7. $S_E = S_E \cup T$. El conjunto de entrenamiento editado se obtiene como el conjunto resultante en S_E .

IV. ESTUDIO EXPERIMENTAL Y VALIDACIÓN DE RESULTADOS

Para el estudio experimental se seleccionaron los 15 conjuntos de datos que a continuación se mencionan: Balance-Scale, Breast-Cancer-Wisconsin, Bupa (Liver Disorders), Dermatology, E-Coli, Heart-Disease (Hungarian), Iris, Lung-Cancer, Monks-1, Pima-Indians-Diabetes, Promoter-Gene-Sequence, Tic-Tac-Toe, Wine Recognition, Yeast y Zoo. Estos son conjuntos de datos internacionalmente reconocidos, para algunos de los cuales ya existen resultados reportados, obtenidos con otros métodos. Los conjuntos de datos son provenientes del depósito de datos para aprendizaje automatizado disponibles en el sitio ftp de la Universidad de Irvine, California¹.

En los experimentos realizados para la comparación de los algoritmos se utilizaron pruebas no paramétricas, tanto para dos muestras relacionadas (prueba de Wilcoxon) como para k muestras relacionadas (prueba de Friedman). Para los casos en que se realizó una prueba de Friedman se aplicó el método de Monte Carlo, con intervalos de confianza del 95% y un número de muestras igual a 10000, mientras que en los casos en los cuales se realizó una prueba de Wilcoxon se aplicó el

¹ <http://www.ics.uci.edu/~mllearn/MLRepository.html>

método de Monte Carlo, con intervalos de confianza del 99% y un número de muestras igual a 10000.

Experimento 1: Comparar los resultados de precisión general y el error cuadrático medio del conjunto de control, obtenidos por los clasificadores k-Vecinos más cercanos (k-NN, IBK del Weka) y la red neuronal Perceptron multicapa (Multilayerperceptron del Weka) para muestras de entrenamiento sin editar, editadas por EditBRS, y editadas por los métodos clásicos de edición: ENN [4], All-kNN [5], Generalized Editing [13] y Multiedit [6].

Se aplicó la prueba de Wilcoxon para los resultados de la precisión general del conjunto de control con el clasificador k-NN, para muestras de entrenamiento sin editar y las editadas por el método propuesto y los resultados arrojaron que las precisiones son significativamente superiores cuando se editan las muestras que cuando se usan muestras sin editar. Se realizó un análisis estadístico similar con el error cuadrático medio y se obtuvo que el error cuadrático medio es significativamente menor con muestras editadas que sin editar. Similar a lo que sucedió con el k-NN, sucedió con el clasificador supervisado Perceptron multicapa.

Se aplicó una prueba de Friedman para las precisiones obtenidas con el conjunto de control para el clasificador k-NN, con muestras de entrenamiento editadas por el EditBRS, ENN, All-kNN, Generalized Editing y Multiedit y arrojó que existen diferencias significativas entre ellas. Se aplicó una prueba de Wilcoxon, dos a dos, y se verificó lo siguiente: Los métodos ENN, All-kNN y Generalized Editing obtienen resultados significativamente inferiores que el método EditBRS, mientras que los resultados de Multiedit no son significativamente diferentes que los del método propuesto.

Lo mismo sucedió con el clasificador MLP. Sin embargo, al realizar el análisis estadístico para el error cuadrático medio, tanto del clasificador k-NN como del Perceptron multicapa, se obtienen resultados significativamente inferiores con muestras editadas por EditBRS que con todas aquellas editadas por los métodos clásicos

Experimento 2: Comparar las muestras editadas por los métodos EditBRS, ENN [4], All-kNN [5], Generalized Editing [13] y Multiedit [6], respecto al porcentaje de reducciones de objetos respecto al conjunto original.

Se aplicó la prueba de Wilcoxon, dos a dos, para las reducciones que se obtienen por los métodos EditBRS, ENN, All-kNN, Generalized Editing y Multiedit y se corrobora que: EditBRS obtienen porcentajes de reducción similares al método Generalized Editing.

Experimento 3: Comparar los resultados de precisión general y precisión por clases del conjunto de control, así como los valores del error cuadrático medio que se obtienen con el clasificador k-NN (IBK del Weka) para muestras de entrenamiento editadas EditBRS y editadas por el filtro de selección de instancias del Weka (Resample).

Se aplicó la prueba de Wilcoxon, dos a dos, y se corroboró lo siguiente: Con las muestras de entrenamiento editadas por EditBRS se obtienen, tanto para la precisión general de la clasificación supervisada como por clases, resultados significativamente superiores que con el método Resample. Al realizar el análisis estadístico respecto al error cuadrático medio se obtuvieron los siguientes resultados: Con muestras editadas por EditBRS se obtienen resultados significativamente menores del error cuadrático medio para el clasificador k-NN, que los obtenidos con las muestras editadas por el método Resample.

Experimento 4: Comparar los resultados de precisión general y precisión por clases del conjunto de control, así como del error cuadrático medio, obtenidos por el clasificador k-NN para muestras de entrenamiento editadas por EditBRS y editadas por los métodos de edición IB1 e IB2 de Aha [8].

Se aplicó la prueba de Wilcoxon, dos a dos, para las precisiones del conjunto de control, tanto general como por clases, del clasificador k-NN para las muestras de entrenamiento editadas por los métodos EditBRS, IB1 e IB2, y se obtienen los siguientes resultados: tanto para el análisis de la precisión general, como por clases, con el método IB1 se logran resultados significativamente inferiores que con el método EditBRS; mientras que no existen diferencias significativas entre el método propuesto y el IB2. Sin embargo, respecto al error cuadrático medio del clasificador k-NN, el EditBRS es superior, pues con las muestras editadas por este, se obtienen resultados del error cuadrático medio significativamente inferiores que los que se obtienen con las muestras que fueron editadas por IB1 e IB2.

La Figura 1 muestra los resultados del desempeño del clasificador k-NN con muestras sin editar y editadas por el método propuesto, para algunos conjuntos de datos seleccionados.

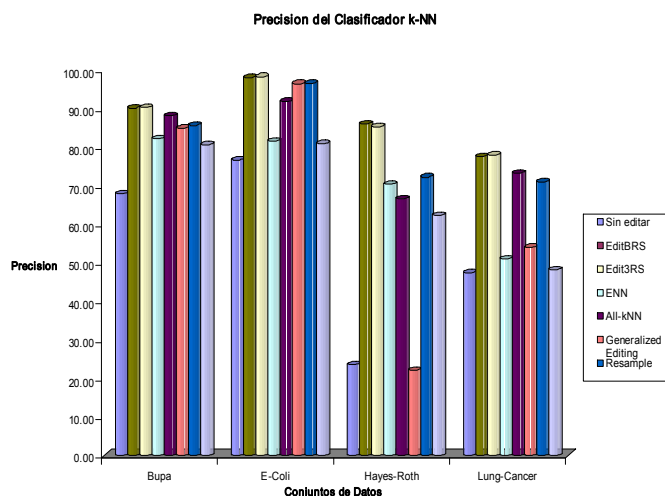


Figura 1. Desempeño del clasificador k-NN

V. VALIDACIÓN EN UN PROBLEMA REAL

El pronóstico de las temperaturas constituye un problema real donde el análisis de datos cobra una alta significación, dada por la complejidad y variabilidad de este proceso a lo largo de los años. En la estación meteorológica de la Universidad Pedagógica y Tecnológica de Colombia, se comenzaron a pronosticar las temperaturas diarias máximas y mínimas a través del sistema SISPROTEM, para estación meteorológica en mención. Este sistema incluye un módulo para la edición de conjuntos de entrenamiento, en el cual se implementa el algoritmo propuesto.

Con los conjuntos de datos iniciales, es decir, sin editar, los resultados de la precisión para el clasificador k-NN no sobrepasaban del 65% de efectividad. Estos conjuntos de datos se editaron por el método propuesto: EditBRS, la tabla 1 muestra los resultados que se obtuvieron respecto a la reducción de la muestra de entrenamiento respecto al conjunto original de datos.

Tabla 1. Resultados de la Reducción obtenida por el método EditBRS, respecto a la muestra original, para la Estación de la U.P.T.C., Colombia.

Base de Casos	% de Reducciones Temperaturas Mínimas	% de Reducciones Temperaturas Máximas
Estación U.P.T.C.	15,02	20,08

En la Tabla 2 se pueden apreciar los resultados promedio que se obtuvieron al tomar como conjunto de control los años 2000 al 2005 y como conjunto de entrenamiento muestras editadas por el método EditBRS, para estación meteorológica de la U.P.T.C.

Tabla 2. Resultados promedio de las precisiones del sistema SISPROTEM para la estación meteorológica de la U.P.T.C., para los años 2000 al 2006

Año	Temperaturas mínimas	Temperaturas máximas
2000	97,94	98,02
2001	98,82	98,97
2002	99,25	99,07
2003	97,47	98,87
2004	98,88	99,09
2005	98,56	98,53
2006	98,03	98,68

En todos los casos estos resultados son significativamente superiores a los obtenidos con las muestras sin editar que ninguna sobrepasó el 78% de efectividad de clasificación.

Las Figuras 2 y 3 muestran el desempeño del pronóstico de las temperaturas con las muestras sin editar y editadas por EditBRS.

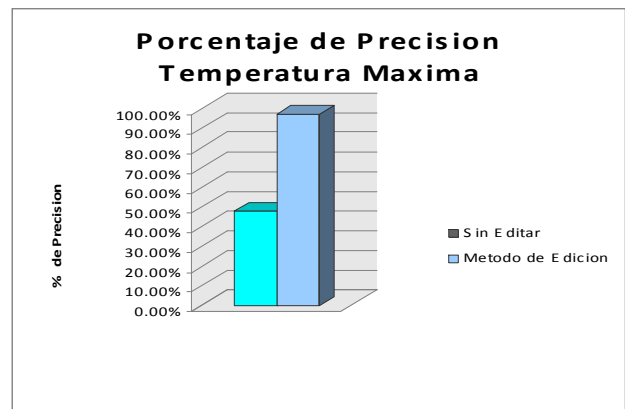


Figura 2. Resultados de precisión para las temperaturas máximas

VI. CONCLUSIONES

En esta investigación se propuso el método EditBRS basado en los conceptos de la Teoría de los Conjuntos Aproximados. Se demostró, a través del estudio experimental, la efectividad de este método en el mejoramiento del desempeño de los clasificadores supervisados, así como la validación en el problema real del pronóstico automatizado de las temperaturas en la estación meteorológica de la U.P.T.C. de Tunja. Este método fue diseñado con el propósito de mejorar la precisión de la clasificación, aunque se logra además, reducir la muestra de entrenamiento respecto a la muestra original.

REFERENCIAS

- [1] García, M., et al., Métodos de selección y construcción de objetos para el mejoramiento de un clasificador supervisado: estado del arte. 2005.
- [2] Hart, P.E., The Condensed Nearest Neighbor Rule. IEEE Transactions on Information Theory, 1968. 14: p. 515-516.
- [3] Gates, G.W., The reduced nearest neighbor rule. IEEE Trans. on Information Theory, 1972. 431-433.
- [4] Wilson, D.L., Asymptotic properties of Nearest Neighbor rules using edited data sets. IEEE Transactions on Systems, Man and Cybernetics, SMC, 1972. 2: p. 408-421.
- [5] Tomek, I., An Experiment with the edited nearest neighbor rule. IEEE Transactions on Systems, Man and Cybernetics, 1976. 6(6): p. 448-452.
- [6] Devijver, P.A. and J.V. Kittler. On the edited nearest neighbor rule. in 5th International Conference on Pattern Recognition. 1980. Los Alamitos, California.
- [7] Kuncheva, L.I., Combining pattern classifiers. Methods and algorithms. 2004: John Wiley & Sons.
- [8] Aha, D., D. Kibler, and A. Marc, InstanceBased Learning Algorithms. Machine Learning. Vol. 6. 1991. 37-66.
- [9] Brighton, H. and C. Mellish, Advances in Instance Selection for Instance-Based Learning Algorithms, in Data Mining and Knowledge Discovery. 2002. p. 53-172.
- [10] Olvera, J.A., J.F. Martínez, and J.A. Carrasco. Edition Schemes Based on BSE. in CIARP 2005. 2005: M. Lazo and A. Sanfeliu (Eds.): LNCS. Springer-Verlag Berlin Heidelberg.
- [11] García, M. and J. Shulcloper, Selecting Prototypes in Mixed Incomplete Data. Lectures Notes in computer Science (LNCS 3773), 2005: p. 450-460.

- [12] Caballero, Y., R. Bello, and Y. Pizano. Improving the k-NN method: Rough Set in edit training set. in *The First IFIP International Conference on Artificial Intelligence in Theory and Practice*. 2006: Springer Boston.
- [13] Koplowitz, J. and T.A. Brown. On the relation of performance to editing in nearest neighbor rule. in *4th International Joint Conference on Pattern Recognition*. 1978. Japan.
- [14] Barandela Ricardo, G.E., Alejo Roberto. , *Correcting the Training Data*. Pattern Recognition and String Matching, 2002.
- [15] Hamamoto, Y., S. Uchimura, and S. Tomita, A bootstrap technique for nearest neighbor classifier design. *IEEE transactions on pattern analysis and Machine intelligence*, 1997. 19: p. 73-79.
- [16] Quenouille, M., Approximation test of correlation in time series,. *Journal of the Royal Statistical Society. Series B* 11 (1949): p. 18–84.
- [17] Bezdek, J.C. and L.I. Kuncheva, *Nearest Prototype classifiers design: an experimental study*. 2004.
- [18] Pawlak, Z., *Rough Sets*. *International journal of Computer and Information Sciences*, 1982. 11: p. 341-356.
- [19] Bosc, P. and H. Prade. An introduction to fuzzy set and possibility theory based approaches to the treatment of uncertainty and imprecision in database management system. in *Proc. of Second Workshop Uncertainty management*. 1993. Information Systems: from Needs to Solution, California.
- [20] Parsons, S., Current approaches to handling imperfect information in data and knowledges bases. *IEEE Trans. On knowledge and data enginnering*, 1996. 8(3).
- [21] Slowinski, R. and D. Vanderpooten, Similarity relation as a basis for rough approximations, in *Advances in Machine Intelligence & Soft-Computing*. 1997. p. 17-33.
- [22] Wilson, D.R. and T.R. Martínez, Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research*, 1997. 6: p. 1-34.
- [23] García, J.M., *KNN Workshop. Suite para el Desarrollo de Clasificadores Basados en Instancias*, in Departamento Computación. Facultad de Matemática, Física y Computación. 2003, Universidad Central “Marta Abreu” de Las Villas.
- [24] Skowron, A. and J. Stepaniuk. Intelligent systems based on rough set approach. in *International Workshop Rough Sets. State of the Art and Perspectives*. 1992.
- [25] Deogun, J.S. Exploiting upper approximations in the rough set methodology. in *First International Conference on Knowledge Discovery and Data Mining*. 1995. Canada.
- [26] Grabowski, A., Basic Properties of Rough Sets and Rough Membership Function. *Journal of Formalized Mathematics*, 2003. 15.
- [27] Efron, B.T., R. J., *Introduction to the bootstrap*. Chapman & Hall, New York,, 1993.