Revista Anales de la Academia de Ciencias de Cuba Vol. 8 No. 1

MÉTODOS PARA LA EDICIÓN Y CLASIFICACIÓN DE CONJUNTOS DE DATOS BALANCEADOS Y NO BALANCEADOS BASADOS EN SOFTCOMPUTING.

ENTIDAD EJECUTORA PRINCIPAL: Universidad de Camagüey[1]

AUTORES PRINCIPALES: DrC. Yailé Caballero Mota[1], DrC. Yaima Filiberto Cabrera[1], DrC. Enislay Ramentol Martínez[1]

OTROS AUTORES: DrC. Rafael Bello Pérez[2], DrC. Nele Verbiest[3], MSc. Mabel Frias Dominguez[1], MSc. Yumilka Fernández Hernández[1], MSc. Yanela Rodríguez Álvarez[1], MSc. Lenniet Coello Blanco[1], DrC. Eduardo Sierra Gil[1], DrC. Chris Cornelis[3], DrC. Francisco Herrera Triguero[3], DrC. Santiago Lajes Choy[1]

COLABORADORES: DrC. Rafael Larrua Quevedo[1], DrC. Israel Gondres Tornes[1], DrC. Sarah Vluymans[3], DrC. Julio Madera Quintana[1], DrC. Ileana Cadenas Freixas[1], DrC. Wilfredo Martinez López del Castillo[1], MSc. Mayte Guerra Saborit[1], MSc. Yanet Sánchez López[1], MSc. Adrian Moreno Garcia[1], MSc. Leander Brizuela Pardo[1], MSc. Yansel Díaz García[1], MSc. Francisco Barrios[4], Ing. Dianne Arias Alvarez[1], Ing. Saily Ojeda Estrada[1], Ing. Erick Machado Alvarez[1], Ing. Rebeca Mulet Deulofeu[1], Ing. Anaira Estévez Batista[1] y Lic. Alina Nordelo Valdivia[5].

OTRAS ENTIDADES PARTICIPANTES:

[2] Universidad Central “Marta Abreu” de Las Villas, [3] Universidad de Granada, España.

[4] Organización Básica Eléctrica Cmg, [5]Centro de Ingeniería Genética y Biotecnología Cmg.

AUTORA PARA LA CORRESPONDENCIA:

DrC. Yaima Filiberto Cabrera.

Circunvalación Norte, entre Camino Viejo de Nuevitas y Ave. Ignacio Agramonte, Camagüey, Cuba. C.P. 74650 Camagüey, Cuba.

e-mail: yaimafiliberto@gmail.com, yaima.filiberto@reduc.edu.cu

RESUMEN:

Los problemas de clasificación aparecen en todas las áreas del conocimiento. Para su solución se utilizan diferentes técnicas, entre ellos se encuentran la clasificación de reglas, la construcción de prototipos y la selección de atributos, todas con el fin de mejorar el rendimiento de los clasificadores. Dada la complejidad de los procesos y las características de la información utilizada para el descubrimiento de conocimiento, donde están presentes diferentes tipos de incertidumbre, se hace necesario el empleo de los métodos de la computación blanda (softcomputing). En esta investigación se aportan nuevos resultados en este contexto; entre ellos se presenta una medida para el cálculo del grado de similaridad basada en la Teoría de los Conjuntos Aproximados (TCA) extendida y en los Conjuntos Borrosos, a partir de la cual se proponen nuevos métodos para realizar el cálculo de los pesos de los atributos y modificaciones al

algoritmo para clasificación de reglas IRBASIR. En el caso de los problemas de construcción de prototipos se presentan dos algoritmos para la generación y selección de los mismos en problemas de clasificación con conjuntos de datos balanceados y no balanceados. Además, se presenta un nuevo algoritmo para la generación de reglas de aprendizaje que utiliza la selección de atributos para obtener el modelo de conocimiento (IRBASIR RED) y se presenta también un nuevo método (REDUCT SIM) para el cálculo de reductos utilizando la técnica de optimización PSO (Particle Swarm Optimization). Se presentan nuevos algoritmos para la clasificación no balanceada usando la TCA, y su combinación con los conjuntos borrosos (TDCA). Los aportes de esta investigación han sido divididos en dos partes: a nivel teórico con los algoritmos propuestos y a nivel práctico con el uso de estos algoritmos para dar solución a problemas reales. La validación de los resultados se ha realizado con bases de datos internacionales y potentes pruebas estadísticas para comparar con los mejores métodos del estado del arte. También, se ha aplicado en la solución de problemas reales en las áreas de la Ingeniería Eléctrica, Ingeniería Civil y la Biotecnología.

Se arriba a las conclusiones siguientes: los métodos propuestos han sido estudiados usando bases de datos internacionales; así como su aplicación para la solución de problemas reales en tres áreas del conocimiento: la Ingeniería Eléctrica, Ingeniería Civil y la Biotecnología. Se han obtenido resultados novedosos y relevantes desde el punto de vista teórico y práctico, lo cual se demuestra en la producción científica asociada y en los 7 avales de introducción de los resultados (3 avales nacionales y 4 internacionales) que se presentan en los anexos.

La producción científica asociada a estos resultados consiste en la publicación de 19 trabajos, de ellos: 17 en revistas y bases de datos referenciadas, 2 libros; así como la presentación de 19 ponencias en prestigiosos eventos científicos internacionales y el registro por CENDA de 5 productos de software. Además, forman parte de los resultados 10 tesis defendidas: 1 tesis de doctorado, 6 tesis de maestría y 3 trabajos de diplomas. Se han obtenido 9 Premios nacionales e internacionales.

Nota: Los resultados que sustentan este trabajo son novedosos y en ningún caso coinciden con los presentados en el Premio de la Academia de Ciencias de Cuba 2013 “Contribuciones al Aprendizaje Automatizado a través de la Teoría de los Conjuntos Aproximados Extendida”, aunque corresponden a la misma línea de investigación. Los aportes merecedores del Premio anterior radicaban en métodos de clasificación y aproximación de funciones, así como métodos de cálculo de pesos en la Teoría de los Conjuntos Aproximados Extendida. Los algoritmos y métodos que se proponen en el presente trabajo se basan en softcomputing, algunos de ellos con enfoque difuso; además, se proponen nuevos métodos de clasificación para conjuntos de datos balanceados y no balanceados y edición basados en prototipos. Las aplicaciones en el 2013 fueron en Ingeniería Civil, fundamentalmente en predicción de estructuras compuestas en bases de datos balanceadas y en Meteorología en el pronóstico meteorológico a mediano y largo plazos; en el presente año se centran en Ingeniería Eléctrica, Biotecnología e Ingeniería Civil (en pronóstico de sucesos en tránsito y en fallo estructural en conjuntos de datos no balanceados).

COMUNICACIÓN CORTA DEL RESULTADO

Los problemas de clasificación aparecen en todas las áreas del conocimiento. Entre ellos se encuentran la clasificación de reglas, la construcción de prototipos y la selección de atributos, todas con el fin de mejorar el rendimiento de los clasificadores. Se presentan como resultados la Medida para determinar el grado de similaridad (RS) basada en la Teoría de los Conjuntos

Aproximados (TCA) extendida y en los Conjuntos Borrosos, la cual es utilizada para realizar el cálculo de los pesos de los atributos y nuevas modificaciones al algoritmo para clasificación de reglas IRBASIR. En el caso de los problemas de construcción de prototipos se presentan dos algoritmos para la generación y selección de los mismos en problemas de clasificación con conjuntos de datos balanceados y no balanceados. Además, se presenta un nuevo algoritmo para la generación de reglas de aprendizaje que utiliza la selección de atributos para obtener el modelo de conocimiento (IRBASIR RED) y se presenta también un nuevo método (REDUCT SIM) para el cálculo de reductos utilizando la técnica de optimización PSO (Particle Swarm Optimization). Se presentan nuevos algoritmos para la clasificación no balanceada usando la TCA, tanto la clásica como la difusa (TDCA). Los aportes de esta investigación han sido divididos en 2 partes: a nivel teórico con los algoritmos propuestos y a nivel práctico con el uso de estos algoritmos para dar solución a problemas reales. La validación de los resultados se ha realizado con conjuntos de bases de datos internacionales valiéndonos de potentes pruebas estadísticas para comparar con los mejores métodos del estado del arte. También, se ha aplicado en el diagnóstico de la necesidad de mantenimiento de interruptores de alta potencia; en la base de casos “Heberprot” pertenecientes al Centro de Ingeniería Genética y Biotecnología de la provincia de Camagüey; en el pronóstico del fallo estructural (si es por el conector o por el hormigón en una estructura compuesta hormigón acero) y el pronóstico del nivel de servicio en vías urbanas.

Los principales resultados de la investigación que constituyen las novedades teóricas y prácticas del trabajo, así como su impacto en la solución de problemas reales son los siguientes: I) Medida calidad de similaridad difusa; II) Método para construir relaciones de similaridad difusa (RSD) y su empleo en clasificadores; III) Nuevos métodos de selección y generación de prototipos basados en la TCA extendida; IV) Métodos de clasificación basados en la medida de similaridad difusa; V) Métodos de clasificación para conjuntos de datos no balanceados; VI) Aplicación de los resultados científicos obtenidos para la predicción de sucesos en la Ingeniería Eléctrica, Ingeniería Civil y Biotecnología.

1. INTRODUCCIÓN

La aplicación de un algoritmo de aprendizaje tiene como objetivo extraer conocimiento de un conjunto de datos y formular dicho conocimiento para su posterior aplicación en la solución de problemas. En el aprendizaje inductivo existen distintas formas de representar el modelo generado, representación proposicional, árboles de decisión, reglas de decisión, listas de decisión, reglas con excepciones, reglas jerárquicas de decisión, reglas difusas y probabilidades, redes neuronales, están entre las estructuras más utilizadas; cuando el conocimiento se maneja de forma implícita se utilizan métodos como el k-NN [1] o razonamiento basado en casos.

Los Conjuntos Borrosos se han investigado ampliamente derivando en exitosas aplicaciones, así los abordan diversos autores: [2, 3-8]. Una de las aplicaciones de la Lógica Borrosa es cuando se quiere construir Conjuntos Borrosos para ser usados en sistemas inteligentes. La TCA es una excelente herramienta matemática para modelar la incertidumbre cuando esta se manifiesta en forma de inconsistencia, permite tratar tanto datos cuantitativos como cualitativos, y no se requiere eliminar las inconsistencias previas al análisis; respecto a la información de salida pues puede ser usada para determinar la relevancia de los atributos, generar las relaciones entre ellos (en forma de reglas), entre otras, [9-12].

El rendimiento de los algoritmos de aprendizaje reales se puede deteriorar ante la abundancia de información. Muchas características pueden ser completamente irrelevantes para el problema (selección de rasgos) y algunos ejemplos pueden constituir ruidos, o generar inconsistencias;

también la cantidad de información puede incrementar el costo computacional del aprendizaje (selección de instancias) [13].

En este último caso los métodos de reducción de datos se dividen en dos enfoques, conocidos como selección de prototipos (SP) [7] y generación/abstracción de prototipos (GP). Estas técnicas permiten obtener un conjunto de entrenamiento representativo con un tamaño menor comparado con el original y obtener un conocimiento con similar o incluso mayor eficacia en la solución de problemas. El primero consiste en seleccionar un subconjunto de los datos de entrenamiento originales, mientras que la GP puede generar nuevos datos artificiales si los necesita. La utilización de estas técnicas en el desarrollo de clasificadores permite obtener nuevos métodos más rápidos, con bajo costo computacional, mejor precisión y más eficientes.

El aprendizaje a partir de datos no balanceados es un problema que afecta el desempeño de los algoritmos de aprendizaje. Sin embargo, y a pesar de existir ya un número considerable de métodos para preprocesar conjuntos no balanceados, éstos no son capaces de generar ejemplos sintéticos y lograr a su vez que pertenezcan con certeza a su clase. Los algoritmos de clasificación tradicionales tienen una debilidad a la hora de enfrentarse a problemas desbalanceados, debido a que tratan las clases mayoritaria y minoritaria de manera simétrica, dejando en desventaja a la menos representada. La Teoría de los conjuntos aproximados, tanto la clásica como la difusa, proveen tres conceptos fundamentales que permiten categorizar los ejemplos en un sistema de decisión atendiendo a si son buenos representantes o no de su clase. Estos conceptos han sido ampliamente utilizados en el preprocesamiento de datos y en sistemas de clasificación [3-5]. En esta investigación se proponen algoritmos que evalúan la calidad de los ejemplos a través de la pertenencia a la región positiva de su clase valiéndose de los dos enfoques de la TCA y usando criterios de edición menos restrictivos para los ejemplos originales; así como diseñar algoritmos de clasificación que analicen la representatividad de las clases con el uso de la agregación con el operador OWA.

2. DESCRIPCIÓN DE LOS RESULTADOS ALCANZADOS.

Todos los resultados de esta investigación incluidos en esta propuesta de premio están definidos para el caso de problemas con datos mezclados, conjuntos de datos no balanceados y balanceados, y se basan en el empleo de los conceptos de la TCA extendida formulados usando relaciones de similaridad y su enfoque difuso. Los aportes principales que se describen en esta sección son una nueva medida denominada calidad de la similaridad borrosa de un sistema de decisión; y un nuevo método para construir relaciones de similaridad borrosa basada en ella. A partir de estos resultados se proponen mejoras al desempeño del método de k-NN [1] y para el entrenamiento de la red neuronal Red Neuronal Multilayer Perceptron MLP [14]; así como nuevos algoritmos de edición y clasificación de conjuntos de datos desbalanceados; nuevos algoritmos de generación de reglas de clasificación tanto para problemas balanceados como con desbalance en los datos, nuevos métodos de edición y clasificación basados en generación y selección de prototipos para conjuntos de datos balanceados y no balanceados, se propone además una estrategia para eliminar ruido en conjuntos no balanceados; y la aplicación de estos nuevos métodos en la implementación de sistemas inteligentes para la predicción en Ingeniería Eléctrica, Ingeniería Civil y Biotecnología. Seguidamente se describen brevemente estos resultados.

Medida calidad de similaridad difusa: Se propone el uso de relaciones borrosas para definir una nueva métrica denominada Medida de la Calidad de la Similaridad Borrosa (MCSB) [15]. La nueva medida utiliza una relación borrosa binaria, que cuantifica la fortaleza de la relación de semejanza entre dos objetos en un rango de [0,1]. La relación borrosa está caracterizada por una función de

pertenencia, que se calcula usando una función de semejanza basada en la Teoría de los Conjuntos Aproximados Extendida.

Método para construir relaciones de similaridad difusa y su empleo en clasificadores: Se propone el método PSO+RST+FUZZY que combina la medida MCSB y PSO [16] para el cálculo de los pesos de los rasgos. Este método permite construir relaciones de similaridad borrosas a partir de funciones de semejanza definidas como una suma pesada de la comparación a nivel de rasgo constituyendo un nuevo procedimiento de aprendizaje de pesos cuya aplicación no está limitada a problemas de clasificación. En este caso, el objetivo del algoritmo de aprendizaje es encontrar el conjunto de pesos, que maximiza MCSB y se estudia su impacto para mejorar el desempeño de los algoritmos k Vecinos más Similares (k-NS) y MLP. Los resultados experimentales en bases de datos internacionales muestran un comportamiento efectivo en cuanto a precisión.

Nuevos métodos de selección y generación de prototipos basados en la TCA extendida: Se proponen tres métodos para la generación y selección de prototipos. Los métodos NPBASIR-CLASS y NPBASIR SEL-CLASS [17] permiten construir y seleccionar los prototipos respectivamente para problemas de clasificación empleando los conceptos de Computación Granular. La granulación de un universo se realiza usando la medida calidad de la similaridad en su definición clásica y borrosa para construir una relación de similitud que genera clases de similitud de objetos en el universo, y para cada clase de similitud se construye/selecciona un prototipo. Otro método de selección de prototipos utiliza la TCA. El mismo elimina del conjunto los ejemplos que no pertenezcan a la región positiva con un umbral δ.

Métodos de clasificación basados en la medida de similaridad difusa: Se propone el empleo de relaciones borrosas en el algoritmo IRBASIR [18, 19] el cual es un método de inducción de reglas de clasificación cuya principal ventaja es su utilización para sistemas de decisión con rasgos de condición heterogéneos, es decir, pueden existir tanto rasgos discretos como continuo. El mismo se distingue de otros métodos en que no requiere discretizar los dominios continuos, y la parte condicional de la regla no se expresa como una conjunción de condiciones elementales. Teniendo en cuenta estas características se proponen tres modificaciones a este algoritmo (IRBASIR-FUZZY) basadas en el uso de las relaciones de similaridad borrosas por las ventajas que estas presentan.

Métodos de clasificación para conjuntos de datos no balanceados: Se proponen los métodos de edición de conjuntos de entrenamiento no balanceados, haciendo uso de operadores genéticos y la TCA, EditOSRS [6] y EditHib2 [7]. Un nuevo método de preprocesamiento basado en la TDCA con doble umbral SMOTE-FRST-2T [8]. Nuevo método de selección de prototipos para conjuntos no balanceados. Mejoras al algoritmo SMOTE para tratar datos con ruido usando la TDCA [20], Uso de vectores de agregación OWA para ponderación de ejemplos en conjuntos no balanceados. Se proponen 6 estrategias para obtener vectores con la agregación OWA [21,22]. Se propone un algoritmo de clasificación IFROWANN para conjuntos no balanceados basado en el método FRNN [23] y la agregación OWA [10]. Un nuevo método de inducción de reglas de clasificación borroso para conjuntos no balanceados, ImbIRBASIR-FUZZY; así como dos nuevos métodos de clasificación basados en prototipos para conjuntos de datos no balanceados.

Aplicación de los resultados científicos obtenidos para la predicción de sucesos en la Ingeniería Eléctrica, Ingeniería Civil y Biotecnología: Se aplicaron los resultados en el diagnóstico de la necesidad de mantenimiento de interruptores de alta potencia [7]. Debido al desbalance observado de los datos captados se aplica el algoritmo SMOTE-FRST-2T, el cual logra magníficos resultados en el uso de la lógica borrosa para la determinación del período de mantenimiento de redes eléctricas, a partir de los datos del. Por otro lado, en el caso de Ingeniería Civil se utiliza el método Imb-IRBASIR [19] por ser un caso de desbalance de clases donde el problema es predecir el modo de fallo estructural, en este caso si es por el conector o por el hormigón en una estructura compuesta hormigón acero; otro problema resuelto es la predicción del nivel de servicio de las vías urbanas en Cuba, utilizando la medida de calidad borrosa con los clasificadores MLP y k-NS, problema de vital importancia en los estudios del tránsito. Se aplican los clasificadores basados en generación y selección de prototipos para el pronóstico del uso del medicamento Heberprot-P en los pacientes que padecen de diabetes. Como ayuda a los especialistas en Ingeniería Eléctrica e Ingeniería Civil se desarrollaron los sistemas automatizados: SIGEMIv1.1, PROCONv4.0, SCBEANv1.0 y Sistema de gestión para el estudio y predicción del tránsito v1.0.

3. CONCLUSIONES

Los métodos propuestos han logrado resultados significativamente superiores a los más reconocidos del estado del arte. Se define una medida, denominada calidad de la similaridad borrosa; usando esta medida se propone un método para construir relaciones de similaridad borrosa. El uso de la RS borrosa y los pesos de los rasgos calculados mediante el método propuesto en varias técnicas del aprendizaje automatizado permitió valorar que estos permiten mejorar el desempeño de estas técnicas. Se han propuesto nuevos algoritmos capaces de balancear conjuntos para luego mejorar el desempeño de los clasificadores en la etapa de aprendizaje. Se han propuesto nuevos algoritmos capaces de aprender directamente de los conjuntos desbalanceados, creando estrategias internas para lidiar la baja representatividad de uno de los conceptos. Las aplicaciones reales abordadas con los métodos propuestos han demostrado la viabilidad y robustez de los mismos. Se solucionaron problemas reales relacionados con las ramas de la Ingeniería Eléctrica, Ingeniería Civil y la Biotecnología

REFERENCIAS

[1] Cover, T.-M. & Hart, P.-E. Nearest neighbour pattern classification Institute of Electronical and Electronics Engineers Transactions on Information Theory, 1967, 13, Pp. 21-27.

[2] Filiberto, Y., et al., Algoritmo para el aprendizaje de reglas de clasificación basado en la teoría de los conjuntos aproximados extendida. Revista DYNA 137, Pp. 16-26, 2011.

[3] E. Ramentol, et al. SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory. International Journal of Knowledge and Information Systems, 33. Pp. 245–265, 2012.

[4] E. Ramentol, N. Verbiest, R. Bello, Y. Caballero, C. Cornelis, F. Herrera. SMOTE-FRST: A new resampling method using FRST. FLINS 2012.

[5] N. Verbiest, C. Cornelis, and R. Jensen. Fuzzy rough positive region-based nearest neighbor classification. FUZZ-IEEE 2012, Pp. 1961–1967, 2012.

[6] E. Ramentol, Y. Sánchez, Y. Caballero, R. Bello F. Herrera. Edición de Conjuntos de Entrenamiento no Balanceados, haciendo uso de Operadores Genéticos y la TCA. Congreso MAEB2009. España.

[7] E. Ramentol, I. Gondres, S. Lajes, Y. Caballero, R. Bello, C. Cornelis, F. Herrera. Fuzzy-Rough Imbalanced Learning for the Diagnosis of High Voltage Circuit Breaker Maintenance: the SMOTE-FTCA-2T Algorithm. Engineering Applications of Artificial Intelligence 48 (2016). Pp. 134–139.

[8] E. Ramentol, et al. IFROWANN: Imbalanced Fuzzy-Rough Ordered Weighted Average Nearest Neighbor Classification. IEEE Transactions on Fuzzy Systems, 2015, Vol. 23, Pp. 1622-1637.

[9] S.K Choubey. A comparison of feature selection algorithms in the context of rough classifiers. In Fifth IEEE International Conference on Fuzzy Systems, 1996.

[10] A. Chouchoulas and Q. Shen. A rough set-based approach to text classification. Lectures Notes in Artificial Intelligence, 11. Pp. 118–127, 1999.

[11] J. W Grzymala-Busse. Managing uncertainty in machine learning from examples. In Proceedings of the Workshop Intelligent Information System III, 1994.

[12] D. Miao and L. Hou. An application of rough sets to monk's problems solving. In 9th International Conference in Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, 2003.

[13] Kohavi, R. and B. Frasca. Useful Feature Subsets and Rough Set Reducts. Third International Workshop on Rough Sets and Soft Computing. 1994.

[14] Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 1958, vol. 65(6), Pp. 386-408.

[15] L. Coello, M. Frias, Y. Fernandez, Y. Filiberto, R. Bello, Y. Caballero. Construction of similarity relations based on the quality of the similarity. Revista Investigacion Operacional, Vol. 38, No. 2, Pp. 132-140, 2017.

[16] Lenniet Coello, Mabel Frias, Yumilka Fernandez, Yaima Filiberto, Rafael Bello, Yailé Caballero. Construcción de relaciones de similaridad borrosa basada en la medida calidad de la similaridad. Special issue "On the inferface between Operations Research and Computational Intelligence: Decision and Optimization Models". REVISTA INVESTIGACION OPERACIONAL, VOL. 38, NO. 2, 132-140, 2017. Scopus.

[17] M. Frias, Y. Fernandez, Y. Filiberto, R. Bello, Y. Caballero. Prototypes Selection Based On Similarity Relations For Classification Problems. En Proceedings of WEA 2015, Colombia. 2015. Pp. 1 – 6, IEEE Press.

[18] Y.B. Fernández-Hernández, et al. An improvement to the classification based on the measurement of the similarity quality using fuzzy relations. DYNA 82 (193), Pp. 70-76. 2015.

[19] Y. Filiberto, et al. "Induction of rules based on similarity relation for imbalance datasets. A case of study. En libro Applied Computer Sciences in Engineering. Vol. 657, No. 1, Pp. 65-73, 414. IEEE Press. 2016

[20] N. Verbiest, et al. Preprocessing Noisy Imbalanced Datasets using SMOTE enhanced with Fuzzy Rough Prototype Selection. International Journal of Applied SoftComputing. (2014). Vol. 22, Pp. 511-517.

[21] C. Cornelis, N. Verbiest, and R. Jensen. Ordered weighted average based fuzzy rough sets. 5th International Conference on Rough Sets and Knowledge Technology, Pp. 78–85, 2010.

[22] N. Verbiest, E. Ramentol, C. Cornelis and F. Herrera. Improving SMOTE with Fuzzy Rough Prototype Selection to detect Noise in Imbalanced Classification data. Iberamia 2012, Vol. 7637, 2012, Pp. 169-178.

[23] R. Jensen and C. Cornelis. Fuzzy rough nearest neighbour classification and prediction. Theoretical Computer Science, 412 (42). Pp. 5871–5884, 2011.

PRODUCCIÓN CIENTÍFICA DE LOS RESULTADOS

La producción científica asociada a estos resultados consiste en la publicación de 19 trabajos, de ellos: 17 en revistas y bases de datos referenciadas, 2 libros; así como la presentación de 19 ponencias en prestigiosos eventos científicos internacionales y el registro por CENDA de 5 productos de software. Además, forman parte de los resultados 10 tesis defendidas: 1 tesis de doctorado, 6 tesis de maestría y 3 trabajos de diplomas. Se han obtenido 9 Premios nacionales e internacionales. Los métodos propuestos han sido estudiados experimentalmente usando bases de datos internacionales; así como su aplicación para la solución de problemas reales en tres áreas del conocimiento: la Ingeniería Eléctrica, Ingeniería Civil y la Biotecnología. Se han obtenido resultados novedosos y relevantes desde el punto de vista teórico y práctico, lo cual se demuestra en la producción científica asociada y en los 7 avales de introducción de los resultados (3 avales nacionales y 4 internacionales) que se presentan en los anexos. A continuación, se detallan los resultados alcanzados. Se han obtenido las siguientes publicaciones científicas en forma de artículos, ponencias y capítulos o epígrafes de libros.

En el Anexo I aparecen los certificados que acreditan estas publicaciones.

Artículos científicos en Revistas y Libros:

1. Yanela Rodríguez Alvarez, Rafael Bello Pérez, Yailé Caballero Mota, Yaima Filiberto Cabrera, Yumilka Fernández Hernández, Mabel Frías Domínguez. A study of the behavior of methods based on prototypes and similarity relations in the face of “hubness”. Revista Cubana de Ciencias Informáticas, Vol. 11, No. 2, Abril-Junio, 2017, ISSN: 2227-1899 | RNPS: 2301 http://rcci.uci.cu, Pág. 134-148. Scielo Citation Index.

2. Lenniet Coello, Mabel Frias, Yumilka Fernandez, Yaima Filiberto, Rafael Bello, Yailé Caballero. Construcción de relaciones de similaridad borrosa basada en la medida calidad de la similaridad. “On the inferface between Operations Research and Computational Intelligence: Decision and Optimization Models”. REVISTA INVESTIGACION OPERACIONAL, VOL. 38, NO. 2, 132-140, 2017. Scopus.

3. Lenniet Coello, Yaima Filiberto, Rafael Bello and Rafael Falcon. Improving the IRBASIR Algorithm with Bayesian Networks. Book Soft computing and Hybrid Systems for Knowledge Discovery and Decision-making as part of Atlantis Computational Intelligence Series (ACIS). Atlantis Press. 2017. Web of Science.

4. Yanela Rodríguez, Rafael Bello, Yailé Caballero, Yaima Filiberto, Yumilka Fernández and Mabel Frias. An Approach to solve Classification Problems on domains with hubness using rough sets and Nearest Prototype. Accepted for publication in a proceedings volume of MICAI 2017 to be published by IEEE CPS. 2017.

5. Dianne Arias, Yaima Filiberto, Rafael Bello, Ileana Cardenas, Wilfredo Martínez. Applied Computer Sciences in Engineering. Classification by Nearest Neighbor and Multilayer Perceptron a New Approach Based on Fuzzy Similarity Quality Measure: A Case Study. Springer International

Publishing AG. Springer International Publishing AG 2017 J.C. Figueroa-García et al. (Eds.): WEA 2017, CCIS 742, pp. 1–10, 2017. DOI: 10.1007/978-3-319-66963-2_35. 2017.

6. Yumilka Bárbara Fernández Hernández, Rafael Bello Pérez, Yaima Filiberto Cabrera, Mabel Frías Dominguez, Yaile Caballero Mota. Efecto de la selección de rasgos en la clasificación basada en prototipos. Revista Cubana de Ciencias Informáticas, Vol. 10, No. 4, Octubre-Diciembre, 2016, ISSN: 2227-1899 | RNPS: 2301, http://rcci.uci.cu, Pág. 83-96, Editorial "Ediciones Futuro", Universidad de las Ciencias Informáticas. La Habana, Cuba, rcci@uci.cu. Scielo Citation Index.

7. E. Ramentol, I. Gondres, S. Lajes, Y. Caballero, R. Bello, C. Cornelis, F. Herrera. Fuzzy-Rough Imbalanced Learning for the Diagnosis of High Voltage Circuit Breaker Maintenance: the SMOTE-FTCA-2T Algorithm. Engineering Applications of Artificial Intelligence 48 (2016) Pp. 134–139. http://dx.doi.org/10.1016/j.engappai.2015.10.009. Sciences Citation Index.

8. Yaima Filiberto, Mabel Frias, Rafael Larrua y Rafael Bello. "Induction of rules based on similarity relation for imbalance datasets. A case of study. 3rd Workshop on Engineering Applications (WEA 2016) held in National University of Colombia-Sede Bogotá from 21-23 September 2016. Book Title Applied Computer Sciences in Engineering, Series Title Communications in Computer and Information Science, Series Volume 657, Publisher Springer International Publishing, eBook ISBN 978-3-319-50880-1, DOI 10.1007/978-3-319-50880-1_6, Softcover ISBN 978-3-319-50879-5, Series ISSN 1865-0929, Edition Number 1, pp 65-73, 414. http://www.springer.com/us/book/9783319508795. Scopus.

9. Coello Blanco, L.; Fernández Hernández, Y.; Filiberto Cabrera, Y.; Caballero, Y.; Bello, R. Impact of weight initialization on multilayer perceptron using Fuzzy Similarity Quality Measure. 3rd Workshop on Engineering Applications (WEA 2016) held in National University of Colombia-Sede Bogotá from 21-23 September 2016. Book Title Applied Computer Sciences in Engineering, Series Title Communications in Computer and Information Science, Series Volume 657, Publisher Springer International Publishing, eBook ISBN 978-3-319-50880-1, DOI 10.1007/978-3-319-50880-1_11, Softcover ISBN 978-3-319-50879-5, Series ISSN 1865-0929, Edition Number 1, pp 115-122, 414. http://www.springer.com/us/book/9783319508795. Scopus.

10. Mabel Frias, Yumilka Fernandez, Yaima Filiberto, Rafael Bello, Yaile Caballero. Prototypes Selection Based On Similarity Relations For Classification Problems. Workshop On Engineering Applications-International Congress On Engineering WEA 2015. PP. 1 – 6, PRINT ISBN: 978-1-5090-0227-6, DOI: 10.1109/WEA.2015.7370130, PUBLISHER: IEEE CONFERENCE # 37658, ATTENDANCE 250. IEEE.UDISTRITAL.EDU.CO/WEA. (Scopus).

11. Yumilka B. Fernández-Hernández, Yaima Filiberto, Mabel Frias, Rafael Bello & Yaile Caballero. An improvement to the classification based on the measurement of the similarity quality using fuzzy relations. DYNA 82 (193), pp. 70-76. October, 2015 Medellín. ISSN 0012-7353 Printed, ISSN 2346-2183 Online DOI: http://dx.doi.org/10.15446/dyna.v82n193.45989. Scopus.

12. Lenniet Coello, Yumilka Fernandez, Yaima Filiberto, Rafael Bello. Improving the Multilayer Perceptron Learning by using a Method to Calculate the Initial Weights with the Quality of Similarity Measure based on Fuzzy Sets and Particle Swarms. Computación y Sistemas, 2015, Vol. 19, No 2 (2015). ISSN 2007-9737 (print version ISSN 1405-5546), Scopus, Thomson Reuters Web of Science.

13. Fernandez, Y.; Bello, R; Filiberto, Y.; Frias, M.; Coello, L.; Caballero, Y. An Approach for Prototype Generation based on Similarity Relations for Problems of Classification. Computación y Sistemas, 2015, Vol 19, No 1 (2015). ISSN 2007-9737 (print version ISSN 1405-5546), Scopus, Thomson Reuters Web of Science.

14. E. Ramentol, S. Vluymans, N. Verviest, Y. Caballero, R. Bello C. Cornelis, F. Herrera. IFROWANN: Imbalanced Fuzzy-Rough Ordered Weighted Average Nearest Neighbor Classification. IEEE Transactions on Fuzzy Systems, 2015, Vol. 23, Pp. 1622-1637. Sciences Citation Index.

15. N. Verbiest, E. Ramentol, C. Cornelis and F. Herrera. Preprocessing Noisy Imbalanced Datasets using SMOTE enhanced with Fuzzy Rough Prototype Selection. International Journal of Applied SoftComputing. (2014). Vol. 22, Pp 511-517. http://dx.doi.org/10.1016/j.asoc.2014.05.023. 1568-4946/© 2014. Sciences Citation Index.

16. Fernandez, Y.; Coello, L; Filiberto, Y.; Bello, R and Falcon, R. Learning Similarity Measures from Data with Fuzzy Sets and Particle Swarms. 2014 11th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE). DOI:10.1109/ICEEE.2014.6978261. ISBN: 978-1-4799-6230-3/14/$31.00 c 2014 IEEE, pp. 296-301. Scopus.

17. E. Sierra, S. Lajes, Y. Filiberto y F. Barrios. Modelo difuso para la determinación del período de mantenimiento de redes eléctricas, a partir de los datos del celaje. Revista DYNA. Revista de la Facultad de Minas de la Universidad Nacional de Colombia. Science Citation Index Expanded (SciSearch), Journal Citation Reports/Science Edition. 2013. Dyna, año 80, Nro. 181, pp. 31-39. Medellín, Octubre, 2013. ISSN 0012-7353. (2013). Scopus.

Libros

1. E. Ramentol, Y. Caballero, R. Bello, F. Herrera. Edición de conjuntos de entrenamiento para datos no balanceados. Libro Tendencias en SoftComputing. Rafael Bello, Raúl Pérez, José Luis Verdegay (Eds), Editorial Feijo, ISBN 959250525-4. 2014.

2. Eduardo Sierra, Santiago Lajes, Yaima Filiberto: Mantenimiento por diagnóstico de redes eléctricas de distribución. Una aproximación desde la lógica difusa. 01/2014; Editorial Académica Española., ISBN: 9783659087899.

OTRAS PUBLICACIONES RECONOCIDAS Y MEMORIAS DE EVENTOS

1. Yanela Rodríguez, Rafael Bello, Yailé Caballero, Yaima Filiberto, Yumilka Fernández and Mabel Frias. An approach for class imbalanced data classification based on Rough Set and Nearest Prototype. ISFUROS 2017. Santa Clara, 2017.

1. Yaima Filiberto, Rafael Bello, Wilfredo Martinez, Dianne Arias and Ileana Cadenas. Prediction by Nearest Neighbor a New Approach based on Fuzzy Similarity Quality Measure. A Case Study. Santa Clara, 2017.

2. Yanela Rodriguez Álvarez, Rafael Bello Pérez, Yaima Filiberto Cabrera, Yaile Caballero Mota, Yumilka Fernández Hernández, Mabel Frias Dominguez. Enfoque basado en prototipos para la clasificación de datos de altas dimensiones. XIV Congreso Nacional de Reconocimiento de Patrones. RECPAT 2016. Santa Clara 2016.

3. Mabel Frias, Yaima Filiberto, Rafael Larrua y Rafael Bello. “Induction of rules based on similarity relation for imbalance datasets. A case of study” and its remarkable presentation in the 3rd Workshop on Engineering Applications (WEA 2016) held in National University of Colombia-Sede Bogotá from 21-23 September 2016. Springer.

4. Coello Blanco, L.; Fernández Hernández, Y.; Filiberto Cabrera, Y.; Caballero, Y.; Bello, R. “Impact of weight initialization on multilayer perceptron using Fuzzy Similarity Quality Measure” and

its remarkable presentation in the 3rd Workshop on Engineering Applications (WEA 2016) held in National University of Colombia-Sede Bogotá from 21-23 September 2016. Springer.

5. Fernández Hernández, Y.; Coello Blanco, L.; Filiberto Cabrera, Y.; Nordelo Valdivia, A.; Bello, R. Una nueva medida de calidad de la similaridad Borrosa utilizada en la predicción de resultados de la aplicación del medicamento heberprot-p en pacientes diabéticos / A new measure of quality of the similarity fuzzy using in the prediction of results in the application of heberprot-p in patient with diabetes. III CONFERENCIA INTERNACIONAL EN CIENCIAS COMPUTACIONALES E INFORMÁTICAS. XVI Convención y Feria Internacional Informática 2016.

6. Frias, Mabel; Filiberto, Yaima; Bello, Rafael; Martínez-López del Castillo, Wilfredo; Cedeño, Osviel. Nuevo modelo de mapa cognitivo difuso con computación con palabras y su aplicación en estudios de vías férreas / New model of fuzzy cognitive map with computing with words and its application in rail road's study. III CONFERENCIA INTERNACIONAL EN CIENCIAS COMPUTACIONALES E INFORMÁTICAS. XVI Convención y Feria Internacional Informática 2016.

7. Lenniet Coello Blanco, Yumilka Fernández Hernández, Yaima Filiberto Cabrera, Rafael Bello Pérez, Julio Madera Quintana. Aplicación de la medida de calidad borrosa en problemas de rendimiento académico. Taller Internacional "La virtualización de la Educación Superior"10 Congreso Internacional de Educación Superior 2016, Habana, Cuba.

8. Mabel Frias, Yumilka Fernández, Yaima Filiberto, Rafael Bello, Yaile Caballero Workshop on Engineering Applications-International Congress on Engineering WEA 2015 de la IEEE. Prototypes selection based on similarity relations for classification problems. Colombia 2015.

9. Lenniet Coello, Yaima Filiberto, Rafael Bello and Rafael Falcon. Fifth international workshop on Knowledge Discovery, Knowledge Management and Decision Support – EUREKA 2015. Improving the IRBASIR Algorithm with Bayesian Networks. México 2015.

10. Fernández, Y.; Coello, L.; Filiberto, Y.; Caballero, Y.; Bello, R. CYTDES 2015. Simposio La Informática por un Desarrollo Sostenible. Aplicación de la Medida Calidad Borrosa en técnicas de clasificación supervisada. Cuba 2015.

11. Fernandez, Y.; Coello, L; Filiberto, Y.; Bello, R and Falcon, R. 11th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE) 2014. Learning Similarity Measures from Data with Fuzzy Sets and Particle Swarms. México 2014.

12. Workshop Rough Sets: Theory & Applications. Granada, España 2014. Using similarity relations in generation of prototypes for problems of classification.

13. XIII Congreso de la Sociedad Cubana de Matemática y Computación COMPUMAT2013. Aumento de la Precisión en el algoritmo IRBASIR, al seleccionar rasgos. Noviembre 2013. La Habana, Cuba

14. E. Ramentol. Nuevas Tendencias en Sistemas Inteligentes y Soft Computing. Granada, España, febrero 2011.

15. E. Ramentol, Y. Caballero, R. Bello. Edición de conjuntos de entrenamiento no balanceado haciendo uso de operadores genéticos y de la Teoría de los Conjuntos Aproximados. Séptimo Congreso de Educación Superior "Universidad 2010". Evento Joven Ciencia.

16. E. Ramentol y colaboradores. Edición de Conjuntos de Entrenamiento no balanceados haciendo uso de Operadores Genéticos y de la Teoría de los Conjuntos Aproximados. VI Congreso Español sobre Metaheurísticas, Algoritmos Evolutivos y Bioinspirados, MAEB´09. ISBN: 978-84-691-6813-4.

17. E. Ramentol, Y. Caballero, R. Bello, F. Herrera. EDITHIB2: Nuevo algoritmo de remuestreo para conjuntos de entrenamiento no balanceados, haciendo uso de operadores genéticos y de la teoría de los conjuntos aproximados. Segundo Taller de Descubrimiento de Conocimiento, Gestión del Conocimiento y Toma de decisiones. Panamá, 4-7 noviembre 2009.

18. Ramentol E., Caballero Y., Bello R. Edición de conjuntos de entrenamiento no balanceados haciendo uso de operadores genéticos y de la teoría de los conjuntos aproximados. Memorias del Evento COMPUMAT2009. ISSN: 1728-6042. 2009

19. E. Ramentol. Seminario Internacional de Soft Computing, Santa Clara 2009.

Autoría de software

Los modelos desarrollados han sido implementados en los siguientes productos de software. En el Anexo II se muestran las respectivas certificaciones.

1. Registro de Software número 3411-11-2016 del Centro Nacional de Derecho de Autor a favor de: Sistema de gestión para el estudio y predicción del tránsito.

2. Registro de Software número 2105-06-2015 del Centro Nacional de Derecho de Autor a favor de: PROCON v4.0. Calcula la capacidad resistente de conectores diversos en estructuras compuestas de acero-hormigón.

3. Registro de Software número 2102-06-2015 del Centro Nacional de Derecho de Autor a favor de: SAICCAD Software. Sistema que combina Selección de Atributos, Inteligencia Colectiva, Conjuntos Aproximados y Difusos.

4. Registro de Software número 2982-09-2014 del Centro Nacional de Derecho de Autor a favor de: SCBEAM v1.0. Programa para el diseño estructural de vigas compuesta de acero y hormigón en situación de incendio v1.0.

5. Registro de Software número 0792-03-2014 del Centro Nacional de Derecho de Autor a favor de: SIGEMI 1.1. Sistema de gestión del mantenimiento de interruptores de alta potencia. 2014.

Tesis de Doctorado en Ciencias Técnicas, Maestrías y Trabajos de Diploma defendidos con éxito a partir de los resultados científicos aquí expuestos.

Tesis de Doctorado:

1. Nuevos métodos de edición de conjuntos de entrenamiento no balanceados usando la Teoría de los Conjuntos Aproximados. Autora: Enislay Ramentol Martínez, de la Universidad de Camagüey, Cuba. Junio, 2014. (Tesis premiada por la Universidad de Granada, España con la Categoría CUM LAUDE).

Tesis de Maestrías:

1. Métodos de aprendizaje basados en prototipos y en relaciones de similaridad: extensiones. 2017. Maestrante: Ing. Yanela Rodríguez Álvarez. Maestría en Informática Aplicada de la Universidad de Camagüey.

2. Métodos para la clasificación en conjuntos de datos no balanceados. 2017. Maestrante: Ing. Mayte Guerra Saborit. Maestría en Informática Aplicada de la Universidad de Camagüey.

3. Métodos de aprendizaje basados en prototipos usando la teoría de los conjuntos aproximados extendida. 2015. Maestrante: Ing. Mabel Frias Domínguez. Maestría en Ciencias de la Computacion de la Universidad Central de Las Villas.

4. Empleo de relaciones de similaridad borrosa para el cálculo de pesos en algoritmos de aprendizaje. 2015. Maestrante: Ing. Lenniet Coello Blanco. Maestría en Ciencias de la Computacion de la Universidad Central de Las Villas.

5. Sistema de apoyo para los especialistas del CDO, utilizando técnicas de inteligencia artificial. 2014. Maestrante: Ing. Yansel Díaz García. Maestría en Informática Aplicada de la Universidad de Camagüey.

6. Sistema de gestión para el ordenamiento territorial. 2014. Maestrante: Ing. Leander Brizuela Pardo. Maestría en Informática Aplicada de la Universidad de Camagüey.

Trabajos de Diploma:

1. Sistema de gestión para el estudio y predicción del tránsito. 2016. Diplomantes: Dianne Áreas Alvarez y Saily Ojeda Estrada. Ingeniería Informática, Universidad de Camagüey.

2. Aplicación de la Inteligencia Artificial a los estudios de Ingeniería de tránsito. Diplomante: Anaira Estévez Batista. Ingeniería Civil, Universidad de Camagüey.

3. Sistema para el aprendizaje de inteligencia colectiva, conjuntos aproximados y difusos. 2015. Diplomante: Rebeca Mulet Deulofeu. Ingeniería Informática, Universidad de Camagüey.

Premios y reconocimientos recibidos.

En el Anexo III aparecen las respectivas certificaciones.

1. **BEST PAPER AWARD** for the outstanding technical quality of the paper entitled “Induction of rules based on similarity relation for imbalance datasets. A case of study” and its remarkable presentation in the **3rd Workshop on Engineering Applications (WEA 2016)** held in National University of Colombia-Sede Bogotá from 21-23 September 2016.

2. **Premio del Rector** al Resultado de mayor impacto científico en las Ciencias Técnicas, en el 2016. Comportamiento termo-estructural y diseño de vigas compuestas de acero y hormigón en situación de incendio.

3. **Premio del Rector** al Mérito Científico Técnico en la categoría: al Colectivo de investigación más destacado en el trabajo de investigación y en la promoción de procesos innovativos en el 2015, con el Grupo Científico de Inteligencia Artificial: AIRES.

4. **Premio del Rector** al Mérito Científico Técnico en la categoría: al Colectivo de investigación más destacado en el trabajo de investigación y en la promoción de procesos innovativos en el 2014, con el Grupo Científico de Inteligencia Artificial: AIRES.

5. **Premio del Rector** al Mérito Científico Técnico en la categoría: al resultado que refleje el avance científico de mayor trascendencia y originalidad durante el año 2014, con el trabajo: Aprendizaje a partir de datos no balanceados usando la Teoría de los Conjuntos Aproximados y su enfoque difuso.

6. **Premio del Rector** al Mérito Científico Técnico en la categoría: al resultado que refleje el avance científico de mayor trascendencia y originalidad durante el año 2014, con el trabajo: Estudio del efecto de la reducción de datos en los métodos de aprendizaje automático usando relaciones de similaridad extendida.

7. **Premio CITMA provincial**. Aprendizaje a partir de datos no balanceados usando la Teoría de los Conjuntos Aproximados y su enfoque difuso. 2014.

8. **Mención** al trabajo tutorado: SCBEAM v1.0. Programa para el diseño estructural de vigas compuestas de acero y hormigón en situación de incendio. **Concurso Nacional de Computación 2014**.

9. Edición de conjuntos de entrenamiento no balanceado haciendo uso de operadores genéticos y de la Teoría de los Conjuntos Aproximados. **Séptimo Congreso de Educación Superior "Universidad 2010"**. Evento Joven Ciencia. **Premio Relevante.**