Pre-procesamiento de conjuntos de entrenamiento de clasificadores del vecino más cercano basado en extensiones a la teoría de los conjuntos aproximados

Autoría principal
Yenny Villuendas Rey[1].

Otros autores
María Matilde García Lorenzo[2], Yailé Caballero Mota[3] y Carmen Rey Benguría[4].

Colaboradores
AA[1],

Entidad ejecutora principal
[1]Facultad de Ciencias informáticas. Universidad "Máximo Gómez Báez" de Ciego de Ávila.

Entidades participantes
[2]Departamento de Computación. Universidad Central "Marta Abreu" de Las Villas.
[3]Departamento de Computación. Universidad de Camagüey.
[4]Centro de Estudios Pedagógicos. Universidad "Máximo Gómez Báez" de Ciego de Ávila.

Autor para correspondencia
Yenny Villuendas Rey.
Facultad de Ciencias Informáticas, Universidad de Ciego de Ávila,
Carretera a Morón, Km 9 ½.
e-mail: yenny@informatica.unica.cu

Aporte científico de cada autor al resultado

- ✓ **Yenny Villuendas Rey** (40%): Es la investigadora principal del proyecto. Coordinó y dirigió los estudios realizados. Recopiló bibliografía de la temática, participó en el desarrollo, redacción y presentación de publicaciones y divulgación de los resultados en eventos científicos. Trabajó en la evaluación, montaje e interpretación de los experimentos.
- ✓ **María Matilde García Lorenzo** (20%): Forma parte de los investigadores del proyecto. Participó en la redacción y presentación de publicaciones y divulgación de los resultados en eventos científicos. Trabajó en la evaluación, montaje e interpretación de los experimentos.
- ✓ **Yailé Cabalero Mota** (20%): Forma parte de los investigadores del proyecto. Participó en la redacción y presentación de publicaciones y divulgación de los resultados en eventos científicos. Trabajó en la evaluación, montaje e interpretación de los experimentos.
- ✓ **Carmen Rey Benguría** (20%): Forma parte de los investigadores del proyecto. Participó en la redacción y presentación de publicaciones y divulgación de los resultados en eventos científicos. Aportó su experiencia para el desarrollo exitoso de las aplicaciones de la investigación a la solución de problemas de la Educación Especial y Preescolar.

Resumen
En esta investigación se enriquece el estado del arte con aportaciones científicas originales, en el ámbito del pre-procesamiento de conjuntos de datos mezclados, incompletos y desbalanceados. Estas se basan en la Teoría de los Conjuntos

Aproximados y en el enfoque Lógico Combinatorio al Reconocimiento de Patrones. Se sustenta un esquema para la selección combinada de rasgos y objetos, con tres variantes generales de algoritmos basados en dicho esquema. Se proponen los Conjuntos Aproximados de Soporte, en el marco de la Teoría Extendida de los Conjuntos Aproximados. Se proponen, además, dos nuevas medidas para evaluar el soporte de un sistema de decisión, las cuales sirven de base para el diseño de tres algoritmos de pre-procesamiento de datos mediante selección de objetos y dos mediante selección de rasgos. Se ha verificado la bondad de los algoritmos propuestos, al probarlos con datos obtenidos de bancos de datos disponibles en repositorios internacionales de gran prestigio. Además, es preciso hacer notar el impacto de los resultados de esta investigación en el ámbito social, en virtud de su exitosa aplicación en el pre-procesamiento de los datos para pronosticar, de forma automatizada, el tipo de orientación que recibirán las familias de menores con trastornos afectivo-conductuales en la escuela “Roberto Ambrosio Zamora Machado” de Ciego de Ávila, y en la detección de niños de edad preescolar con altas potencialidades para el desarrollo, también en la provincia de Ciego de Ávila. Por último, se considera que las medidas propuestas son confiables, y que existen evidencias acerca de su validez.

Comunicación Corta

Resumen

El mejoramiento de clasificadores supervisados es esencial en la solución de problemas de clasificación. En este sentido, el pre-procesamiento del conjunto de entrenamiento juega un papel muy importante. En este trabajo se proponen los Conjuntos Aproximados de Soporte. Esta propuesta sirve de sustento teórico al desarrollo de un esquema de pre-procesamiento de datos, con tres variantes generales de algoritmos basados en dicho esquema. Se proponen, además, dos nuevas medidas para evaluar el soporte de un sistema de decisión, las cuales sirven de base para el diseño de tres algoritmos de pre-procesamiento de datos mediante selección de objetos y dos mediante selección de rasgos. La bondad de las propuestas realizadas fue evaluada en bases de datos internacionales, evidenciándose sus buenos resultados. Por otra parte, es preciso hacer notar el impacto de los resultados de esta investigación en el ámbito social, en virtud de su exitosa aplicación para pronosticar, de forma automatizada, el tipo de orientación que recibirán las familias de menores con trastornos afectivo-conductuales en la escuela “Roberto Ambrosio Zamora Machado” de Ciego de Ávila, y en la detección de niños de edad preescolar con altas potencialidades para el desarrollo, también en la provincia de Ciego de Ávila.

Introducción

La selección de los objetos y los rasgos que formarán el conjunto de entrenamiento es fundamental para mejorar la eficiencia y eficacia de muchos clasificadores supervisados, particularmente de la regla del vecino más cercano (NN). Aunque se han propuesto algoritmos de pre-procesamiento de datos, estos presentan limitaciones para su aplicación a la solución de problemas del ámbito social, debido fundamentalmente a su alto componente estocástico, pobre manejo de ruido, incapacidad de atacar problemas con datos mezclados e incompletos, y elevado costo computacional. Es por

ello que en esta investigación se propone un nuevo esquema para resolver esta problemática.

1. Esquema para el pre-procesamiento de conjuntos de entrenamiento.

El esquema que se propone, denominado IFIS (Intelligent Feature and Instance Selection) [1, 2], funciona en cuatro etapas: pre-procesamiento de objetos, obtención del sistema de conjuntos de apoyo, creación de submatrices y fusión de submatrices.

A continuación se explican estas etapas. La fase de pre-procesamiento de objetos de IFIS, que le otorga el nombre de "inteligente", consiste en, dadas las características de solapamiento y desbalance del conjunto de entrenamiento, filtrar este mediante la selección de objetos, ya sea por edición [3, 4], condensación [5] o balanceo de clases [6].

El esquema IFIS no impone ningún algoritmo en particular para la obtención de un sistema de conjuntos de apoyo, pero sí necesita de alguna vía para obtener este. Ejemplos de algoritmos para la obtención de sistemas de conjuntos de apoyo son el algoritmo BMFS [2] que permite la obtención de múltiples conjuntos de rasgos, o los algoritmos GAMFS y BCMFS [7]. Cabe señalar que en el esquema IFIS, la obtención del sistema de conjuntos de apoyo se realiza paralelamente al pre-procesamiento del conjunto de entrenamiento; es decir, utilizando el conjunto de entrenamiento original, no el pre-procesado. El esquema IFIS, para la obtención de las sub-matrices, proyecta el conjunto de entrenamiento pre-procesado, usando cada uno de los conjuntos de apoyo, y a dicha proyección le aplica un método de selección de objetos.

En esta investigación se propone la combinación de elementos del enfoque Lógico Combinatorio al Reconocimiento de Patrones y de la Teoría Extendida de los Conjuntos Aproximados, en la definición de los Conjuntos Aproximados de Soporte (Support Rough Sets) [5]. Estos conjuntos se utilizaron en el desarrollo de los métodos para la selección de objetos [3, 4, 6] y para los métodos de selección de múltiples conjuntos de rasgos [2, 7]. En estos últimos, se aplicaron las medidas de soporte obtenidas como parte de esta investigación [5]. Se introdujeron además 3 variantes de algoritmos de selección de rasgos y objetos, que utilizan elementos de la computación bio-inspirada, y de combinación de múltiples clasificadores. Ellos son los algoritmos IFIS-G [1, 8], IFIS-ANT [9] e IFIS-SA [10].

Resultados Experimentales obtenidos con los algoritmos propuestos, en bases de datos internacionales

En esta investigación se decidió medir la eficacia del clasificador teniendo en cuenta el promedio de su exactitud por clase. Se usaron bases de datos del repositorio de Aprendizaje Automático de la Universidad de California en Irvine [11]. En la evaluación del desempeño de los algoritmos, se utilizó el procedimiento de validación cruzada en 10 hojas, excepto para el caso no balanceado, donde se utilizaron cinco hojas, siguiendo las sugerencias de [12]. Los algoritmos con componente estocástico fueron evaluados en 10 ocasiones diferentes, y se promediaron sus resultados.

Para determinar la existencia o no de diferencias entre los algoritmos comparados, se utilizaron pruebas no paramétricas. Se utilizó el test de Iman y Davenport para la comparación de varias muestras relacionadas, y el test de Holm para dos muestras relacionadas. Estos test son sugeridos en [13], donde además se ofrece una herramienta para su cálculo. Los resultados experimentales muestran que el esquema de edición combinada de rasgos y objetos favorece los resultados que se alcanzan con los clasificadores basados en los vecinos más cercanos ya que muestran buenos resultados tanto en desempeño del clasificador, como en retención de rasgos y objetos. Además, los algoritmos de selección de rasgos y objetos pertenecientes al esquema propuesto superan a otros de su tipo, puesto que tienen mejores resultados con al menos una de las medidas de desempeño a evaluar e iguales resultados con el resto de las medidas.

Resultados Experimentales obtenidos con los algoritmos propuestos, en problemas reales

El proceso de orientación a familias de menores con trastornos de la conducta en la provincia de Ciego de Ávila, fue uno de los problemas a resolver por medio de las propuestas presentadas en esta investigación [14, 15]. En particular, cabe destacar que se obtuvo un 100% de clasificaciones correctas, utilizando un número reducido de rasgos y objetos. Del análisis de los resultados obtenidos, se puede concluir que varios rasgos no eran necesarios para el proceso de clasificación, por lo cual los especialistas de la escuela de conducta ya no necesitan determinarlos. Los rasgos que sí resultaron necesarios para la clasificación son: impacto, actitud, culpa, manejo, relaciones, impotencia, conciencia y tiempo. Por otra parte, se obtuvo un conjunto de familias “típicas”, que pueden ser utilizadas exitosamente como representantes de los grupos de orientación, para la comparación con las nuevas familias de menores que arriben a la escuela de conducta. Estos resultados permitieron a los especialistas de la escuela, diseñar un sistema personalizado de orientación familiar, acorde con las peculiaridades de cada clase de familias, y ubicar a las nuevas familias en el grupo más afín con sus características, para dar una mejor respuesta a sus necesidades.

Por otra parte, la detección temprana de niños del grado preescolar con altas potencialidades para el desarrollo también fue un problema de la práctica pedagógica de la provincia de Ciego de Ávila, que fue impactado positivamente con los resultados de la presente investigación. En este caso, los algoritmos pertenecientes al esquema propuesto, obtuvieron buenos resultados, utilizando un número reducido de rasgos y objetos. El algoritmo con mayor eficacia fue el IFIS-SA, que se diferenció en solo 0.06 con el clasificador original, utilizando todos los rasgos y objetos [7]. Del análisis de los resultados obtenidos, se puedo concluir que varios rasgos no eran necesarios para el proceso de detección de la alta potencialidad, por lo cual no es necesario determinarlos. Los rasgos que no se consideraron necesarios fueron: edad, familia, actividad, autoestima, rapidez, juego, precedente y desempeño.

Cabe señalar que esta investigación se encuentra en curso en el municipio de Ciego de Ávila, al momento de la escritura de esta propuesta, y que cuando se cuente con los datos de todos los niños del grado preescolar del municipio, los resultados preliminares obtenidos en esta investigación pueden estar sujetos a cambios. Igualmente, para la

generalización de estos resultados a otros municipios y provincias se debe tener en cuenta que las características de los niños pueden variar, y se sugiere aplicar los resultados de esta investigación en varios lugares diferentes, antes de tomar decisiones concernientes al currículo de la Enseñanza Preescolar.

Conclusiones

El método CSE propuesto trabaja con MID y con similaridades más generales (incluso no simétricas o no definidas positivas). Produce un conjunto consistente según las subclases. Se mostró que el algoritmo propuesto tiene un buen desempeño comparado con otros que también pueden ser usados para MID después de extensiones triviales. El nuevo método no es ni un método de condensación puro ni un método de edición puro, sino que tiene características deseables de ambos. Basado en experimentos preliminares y los resultados alcanzados, CSE resulta un método adecuado para sinergia de métodos de selección de objetos.

Referencias

[1] Villuendas-Rey, Y., Caballero-Mota, Y., García-Lorenzo, M.M.: Intelligent feature and instance selection to improve Nearest Neighbors classifiers. 11th Mexican International Conference on Artificial Intelligence, MICAI 2012. Springer-Verlag (2012)

[2] Villuendas-Rey, Y., García-Lorenzo, M.M.: Attributes and cases selection for NN classifier through rough sets and natural inspired algorithms. Computación y Sistemas 18 (2014) 293-309

[3] Villuendas-Rey, Y., Caballero-Mota, Y., García-Lorenzo, M.M.: Using Rough Sets and Maximum Similarity Graphs for Nearest Prototype Classification. Lecture Notes on Computer Science 7441 (2012) 300-307

[4] Villuendas-Rey, Y., Caballero-Mota, Y., García-Lorenzo, M.M.: Prototype selection with Compact Sets and Rough Sets. Lecture Notes on Artificial Intelligence 7637 (2012) 159-168

[5] Villuendas-Rey, Y., García-Lorenzo, M.M., Bello-Pérez, R.: Support Rough Set for Decision-making. Eureka 2013. Atlantis Press, México (2013)

[6] Villuendas-Rey, Y., Garcia-Lorenzo, M.M.: Mixed Data Balancing through Compact Sets Based Instance Selection. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer (2013) 254-261

[7] Villuendas-Rey, Y.: Esquema para el Pre-procesamiento de conjuntos de entrenamiento de clasificadores del vecino más cercano basado en extensiones a la teoría de los conjuntos aproximados. Departamento de Computación, Vol. Tesis Doctoral. Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Villa Clara (2014)

[8] Villuendas-Rey, Y., Caballero-Mota, Y., García-Lorenzo, M.M.: Improving case based decision making by hybridizing Rough Set Theory and Swarm Intelligence. II Conferencia Internacional de Ciencias Computacionales e Informáticas, CICCI 2013, La Habana, Cuba (2013)

[9] Cabrera-Cano, Y., Villuendas-Rey, Y., Caballero-Mota, Y., García-Lorenzo, M.M.: Training set preprocessing through swarm intelligence algorithms. International Congress COMPUMAT 2013 (in spanish). Cuban Society of Mathematics and Computers, La Habana, Cuba (2013)

[10] Rigondeaux Caraballo, A.: Pre-procesamiento de conjuntos de entrenamiento a través de la selección combinada de rasgos y objetos. Departamento de Ciencias Informáticas, Vol. Trabajo de Diploma en Opción al Título de Ingeniero en Informática. Universidad "Máximo Gómez Báez" de Ciego de Ávila, Ciego de Ávila (2013)

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

[12] García, S., Derrac, J., Triguero, I., Carmona, C.J., Herrera, F.: Evolutionary-based selection of generalized instances for imbalanced classification. Knowledge Based Systems 25 (2012) 3-12

[13] García, S., Herrera, F.: An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. Journal of Machine Learning Research 9 (2008) 2677-2694

[14] Villuendas-Rey, Y., Caballero-Mota, Y., García-Lorenzo, M.M.: Nearest prototype classification of special school families based on hierarchical compact sets clustering. Lecture Notes on Artificial Intelligence 7637 (2012) 662-671

[15] Villuendas-Rey, Y., Rey-Benguría, C.F., Caballero-Mota, Y., García-Lorenzo, M.M.: Improving the family orientation process in Cuban Special Schools through Nearest Prototype Classification. International Journal of Artificial Intelligence and Interactive Multimedia, special Issue on Artificial Intelligence and Social Application 2 (2013) 12-22