Clasificación automática de objetos utilizando sistemas inteligentes

Fernando Lage [(1)], Zulma Cataldi [(1)], Gregorio Perichinsky [(2)] y Ramón García–Martínez [(3)]

(1) LIEMA - Laboratorio de Informática Educativa y Medios Audiovisuales
(2) LBDySO - Laboratorio de Bases de Datos y Sistemas Operativos
(3) LSI - Laboratorio de Sistemas Inteligentes Facultad de Ingeniería. UBA y Centro de Ingeniería del Software e Ingeniería del Conocimiento. Escuela de Postgrado. ITBA.

Grupo de I+D en Ingeniería Informática. Facultad de Ingeniería. UBA. liema@fi.uba.ar

Resumen

Esta línea de investigación se centra en la aplicabilidad de los Sistemas Inteligentes (SI) a la clasificación automática de objetos a fin de solucionar algunos de los problemas tales como los límites difusos en sistemas diversos.
Se la presenta como una continuación de los desarrollos efectuados por Perichinsky et al. [1989,2000] en el Laboratorio de Bases de Datos y Sistemas Operativos, tendiente a obtener taxonomías o clasificaciones a través de *clusters* en sistemas de diferentes tipos.
Tal es el caso de los asteroides como caso de estudio [Perichinsky, 2000], ya que la clasificación manual es lenta, imprecisa y laboriosa debido a las estimaciones que requieren tiempo y experticia del clasificador humano. Es por ello, que se requiere de nuevos métodos que permitan clasificar con eficiencia a través del reconocimiento de las agrupaciones, objetivo que podría alcanzarse incorporando sistemas orientados al autoaprendizaje [García Martínez, 1996] o metaaprendizaje.

Palabras Clave: clustering, clasificación automática.

Introducción

Se ha observado que la mayoría de los sistemas naturales (ya sean biológicos, físicos, químicos u otros) presentan como característica general que los límites entre distintas clases son muy difusos. Esta cuestión, ha dificultado a los especialistas la tarea al intentar definir en qué grupo se deben colocar aquellas instancias que se hallan en la frontera entre dos o más clases disjuntas de un todo. [Gennari, 1989]

La clasificación es una técnica de abstracción utilizada para agrupar objetos con propiedades comunes. Se espera que tal clasificación delimite el dominio de los objetos, con *la hipótesis de que cada objeto pertenece a una (y solo una) clase y que, para cada clase, hay al menos n objetos que pertenece a ella* [Perichinsky et al., 2002].
Uno de los problemas críticos a resolver cuando se crea una taxonomía sobre un conjunto dado es la búsqueda de conceptos clasificatorios que permitan una estructura de clasificación que no se modifique por el agregado de nueva información (estabilidad de la clasificación), ni se altere a través de la incorporación de nuevas entidades. [Perichinsky et al., 2000, 2003].

Por este motivo, se ha pensado que una de las soluciones factibles a este problema podría encontrarse en la aplicación de los SI a fin de disminuir los problemas de límites en la clasificación de los distintos objetos. De este modo, se evitarían esfuerzos innecesarios en una primera etapa y en la medida el sistema inteligente seleccionado evolucionara hacia el autoaprendizaje [García Martínez, 1996] podría ir redefiniendo sus propios patrones de clasificación. Esta solución podría aplicarse a datos almacenados en bases de datos con dominios dinámicos, ya que cada vez que el sistema redefiniera los patrones de clasificación, se podría comprobar la ubicación de los objetos anteriormente

clasificados, no permitiendo, de este modo incongruencias entre los datos almacenados en la base. [Perichinsky, 1989]

Los SI permitirían obtener las mismas familias que a través de otros métodos clasificatorios, pero se piensa que el desarrollo de un sistema de software sería una herramienta metodológica que permitiría sustentar las aplicaciones para cualquier disciplina a fin de realizar estudios de grupos estables sin importar los problemas de la escalabilidad en los resultados y las propiedades de los espacios que se estudien, ya que en taxonomía los hiperespacios son homogéneos e isótropos. (Perichinsky et al., 2000, 2003)

Existen trabajos clasificación automática orientados a tópicos específicos como la clasificación automática de documentos aplicando algoritmos genéticos. En particular, el *análisis de cluster* ha sido estudiado en estadística y en biología, siendo muy similar al estudio de *formación de conceptos* en aprendizaje automático. En este sentido, se considera que los estudios acerca del análisis de clusters pueden orientar las investigaciones en aprendizaje automático en esa área con las restricciones de la atribución de significación como carácter inherente de los seres humanos. [Pozo, 1994]

Pero, en el sentido propuesto, no se han encontrado trabajos previos orientados a este tipo de aplicaciones con vistas a la generalización de las soluciones.

Descripción

El método de clasificación comúnmente utilizado se denomina taxonomía numérica. La *taxonomía numérica* es el estudio teórico de la clasificación que incluye las bases los principios, los procedimientos y las reglas. Es el producto de la clasificación numérica o proceso taxonómico [Sneath, 1973]

Su propósito es el de obtener teórica y prácticamente procesos clasificatorios, contrastando la visión convencional con conceptos susceptibles de evolución. [García Martínez, 1996]

Cuando se habla de *"Clustering",* se hace referencia a la clasificación de un conjunto de instancias en clases (clusters), siendo estas instancias descriptas por pares atributo-valor; se busca el conjunto de clases que agrupe estas instancias en función de estos pares de atributos.[Perichinsky, 2003].

Para esta tarea de clasificación se requiere una *"función de evaluación",* a fin de determinar a que clase pertenece una cierta instancia. En la mayoría de los casos la definición de esta función se constituye en una dificultad central.

Esta función de evaluación está fuertemente ligada a los criterios de similitud que determinan las instancias, es decir a los atributos que las descubren.

La importancia del tema radica en que una de la aplicaciones de estos sistemas son las Bases de Datos con Dominios Dinámicos (BDDD) [Perichinsky, año] que permite obtener un nuevo enfoque en la investigaciones sobre Bases de Datos [de Miguel,2000], tradicionales (BD) que se centran en el almacenamiento de los datos por única vez, con independencia de su tratamiento.
Este enfoque permite obtener un modelo de Base de Datos (BD) en el cual se trata a los dominios en forma independiente.
En este marco, los Sistemas Inteligentes emergen como un campo disciplinar que permite el estudio y el desarrollo de algoritmos que facilitan la implementación de las diferentes formas de aprendizaje (tales como el aprendizaje por analogía, a través de ejemplos, por inducción fundamentada en datos, etc) tratando de buscar la solución a problemas prácticos a través de su aplicación. [Michalski, 1983; Dejong y Money 1986; Bergadano et al., 1992]. De este modo se pueden resolver problemas de difícil solución y otros imposibles de solucionar utilizando métodos algorítmicos tradicionales.
Como ejemplos se pueden enumerar desde los casos de establecimiento de condiciones asociadas a diagnósticos técnicos o clínicos, la identificación de características para el reconocimiento visual de formas y de objetos, el descubrimiento de patrones o regularidades en estructuras de información

tales como en bases de datos de gran tamaño, [Perichinsky, 2003] que podrían resolver problemas asociados a áreas diversas como la búsqueda a través de las bases de datos agronómicos, hasta la búsqueda a través de bases con datos clínicos que podrían facilitar soluciones a diferentes grupos con riesgos diversos.
Dentro de los SI, en estudio, se encuentran las redes neuronales (RN), que se pueden definir como elementos simples que se hallan interconectados masivamente en paralelo, con una organización jerárquica, y que intentan interactuar con los objetos del mundo real a modo de un sistema neuronal psicológico [Kohonen, 1988].
Las redes operan en forma más compleja cuando se consideran de las condiciones de retroalimentación a través de los lazos en los procesos no lineales entre sus elementos, teniendo además, cambios no adaptativos de sus parámetros que pueden descubrir fenómenos de manera muy complicada. [Hilera González y Martínez Hernando, 2000]
Una característica importante de las RN es que pueden aprender de la experiencia generalizando a través de la generalización de casos, es decir abstrayendo las características esenciales a través de los datos de entrada.

Para el caso de algoritmos genéticos (AG) se estima que los mismos pueden contribuir a la resolución del problema de las clasificación automática de los objetos limítrofes [Raghavan y Birchard,1978; Johnson y Fotoui, 1996, Cole, 1998, Paino y Bação, 2000], ya que se caracterizan por su capacidad de explorar el espacio de búsqueda en forma amplia y eficiente [Goldberg, 1989, Koza, 1997]
Los algoritmos genéticos son una abstracción del concepto biológico de evolución natural y se los puede aplicar en problemas de optimización [Davis, 1991; Falkenauer, 1999]. Su funcionamiento se basa en los mecanismos de la selección natural, a través de la supervivencia del más apto en un intercambio de información entre miembros de una población de posibles soluciones.
En este contexto, se busca indagar cómo se pueden aplicar los AG para resolver el problema de la clasificación automática con límites difusos y estimar la calidad de la solución obtenida.

El enfoque basado en agentes inteligentes (AI), considera que la inteligencia genuina sólo es posible si se cuenta con un cuerpo situado dentro de un entorno (actuador), aspecto totalmente descuidado en los anteriores paradigmas. Para interactuar con el medio ambiente, en consecuencia, el agente debe ser capaz de *percibir, razonar y actuar*. En este sentido, debe poseer algún tipo de sensores que le permitan recoger información como percepciones; debe ser capaz de convertir de algún modo esa información en conocimiento para poder utilizarlo luego para alcanzar sus objetivos es decir razonar. Finalmente debe disponer de algún tipo de actuador a fin de poder modificar el entorno. Por este motivo, podría se podría decir que este enfoque se basada en el modelo biológico y se habla entonces de conceptos tales como “deseos”, “motivaciones”, “creencias” y “humores”. Si se considera un grupo de agentes cooperando (y por ende compitiendo) entre sí, se está dentro de los que se podría denominar sistemas ecológicos [Huberman, 2000] y se habla de “comunicación”, “cooperación”, “coordinación” y “competencia”. Por esta razón, se podría afirmar que este enfoque se inspira en la sociología. [Moriello, 2003]

Objetivos

Los objetivos generales planteados a largo plazo para esta línea que considera la inclusión de los SI en los métodos tradicionales se centran en:

– *El estudio de factibilidad de implementación de los diferentes SI para mejorar la eficiencia algorítmica de los métodos de clasificación automática convencionales.*

– *La construcción de un ambiente para efectuar los estudios comparativos y las evaluaciones de las diferentes opciones.*

Cada uno de los anteriormente enunciados como objetivos generales puede desglosarse en la siguiente serie de objetivos específicos que permitirán su concreción. Por ejemplo el primero de ellos lo hará en:

- *Analizar los diferentes tipos de SI de posible aplicación.*
- *Efectuar un estudio descriptivo y de factibilidad de para uno de los posibles SI a utilizar.*

Por otra parte, se estima una metodología de trabajo que combina los métodos, las técnicas y las herramientas básicas utilizadas en ingeniería de software con aquellas que proveen los diferentes tipos de sistemas inteligentes. En este sentido la articulación de las soluciones se orientará hacia la búsqueda de algoritmos más eficientes en un ambiente que permita efectuar estudios comparativos y evaluaciones de las diferentes soluciones. Básicamente se busca la generalización de los resultados a casos de clasificación en diferentes ámbitos (físicos, biológicos, agronómicos, astronómicos, etc.) centrados en la misma problemática de agrupamientos con límites difusos.

Referencias bibliográficas

Bergadano et al., (1999) A Java-Based and Secure Learning Agents Architecture for Information Retrieval in Distributed Systems. Information Sciences, (113):55--84, 1999.

Cole, R. M. (1998). Clustering with Genetic Algorithms. Thesis for the degree of Master of Science, Department of Computer Science, University of Western Australia.

Davis, L. (1991). Handbook of Genetic Algorithms. New York. Van Nostrand Reihold.

de Miguel Castaño A,(2001) Diseño de Bases de Datos, Editorial ALFAOMEGA España (PP 456-494)

De Jong, K A. (1980). Adaptive system design: A genetic approach. IEEE Transactions on Systems, Man, and Cybernetics, SMC - 10(9), 566-574

Falkenauer, E. (1999). Evolutionary Algorithms: Applying Genetic Algorithms to Real-World Problems. Springer, New York, Pag 65--88.

Gallant, S. (1993) Neural Network Learning and Experts Systems. MIT Press, Cambridge, MA.

García Martínez R. (1997) Sistemas Autónomos. Aprendizaje Automático. Nueva Librería.

García Martínez, R. (1994). Adquisición de Conocimiento. En Abecasis, S. y Heras, C. Metodología de la Investigación. Prologado por el Dr. L. Santaló. 157 páginas. Editorial Nueva Librería. ISBN 950-9088-65-x.

García Martínez, R. et al.(1996) Unsupervised Machine Learning Embedded in Autonomous Intelligent Systems. Proceedings of the XIV International Conference on Applied Informatics. Páginas 71-73. Innsbruck. Austria. 1996.

García Martínez, R. et al (2000) An Integrated Approach of Learning, Planning and Executing. Journal of Intelligent and Robotic Systems. Volumen 29, Número 1, Páginas 47-78. Kluwer Academic Press. 2000

Gennari, J. H. (1989) A Survey of Clustering Methods, Departamento de Informatica y Ciencias de la Computación, Universidad de California., Irvine, CA 92717, Páginas 4-5

Goldberg, D. E. (1989). Genetic Algorithms - in Search, Optimization and Machine Learning. Addison-Wesley Publishing Company, Inc.

Hilera González y Martínez Hernando, 2000] Redes Neuronales Artificiales: Fundamentos, modelos y aplicaciones. Ra-ma, Madrid.

Johnson y Fotoui, (1996). Adaptive Clustering of Hypermedia Documents. Information Systems, Vol 21 Nª 6, (pp 469)

Kohonen, T. (1988) Self-Organizing Maps Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, NY(pp 236)

Koza, John R. (1997). Genetic Programming. Cambridge : M.I.T. Press.

Michalski, R. S. (1991). Toward an Unified Theory of Learning: An Outline of Basic Ideas, Proceedings of the 3rd World Conference on the Fundamentals of Artificial Intelligence, Paris, Julio 1-5, 1991

Michalski, R. S., A (1983) Theory and Methodology of Inductive Learning, Artificial Intelligence, 20:111-161,

Michalski, R. S., Carbonell, J. G., Mitchell, T. M. (eds.), (1986). Machine Learning: An Artificial Intelligence Approach, Vol. II, Morgan-Kauffman

Michalski, R. S., Carbonell, J. G., Mitchell, T. M. (eds.). (1983). Machine Learning: An Artificial Intelligence Approach, Vol. I. Morgan-Kauffman

Michalski, R. S., Tecuci, G. (eds) (1994). Machine Learning: A Multistrategy Approach, Vol. IV, Morgan Kauffman

Michie, D. (1988). Machine Learning in the next five years, EWSL-88, 3rd European Working Session on Learning, Glasgow, Londres, Pitman.

Mitchell, T. M. (1996). Machine Learning, McGraw-Hill.

Moriello, S. (2003) Tesis de Magíster en agentes inteligentes. UTN-FRBA.

Painho, M. y Bação, F. (2000). Using Genetic Algorithms in Clustering Problems. Proceedings of the 5th International Conference on GeoComputation, University of Greenwich, United Kingdom.

Perichinsky, G. et al. (1989). Base de datos relacional estructurada sobre dominios dinámicos de atributos. Exposición en las 18 JAIIO. Buenos Aires.

Perichinsky, G et al. (2000). Knowledge Discovery Based on Computational Taxonomy and Intelligent Data Mining. VI Congreso Argentino de Ciencias de la Computación, CACIC, CD. Universidad Nacional San Juan Bosco. Sede Ushuaia. Argentina.

Perichinsky, G et al. (2002). Spectra of Taxonomic Evidence in Databases.III. Application in Celestial Bodies. Asteroids families. Pag. 212-226. International Association for (ACIS) Conference on Computer Science, Software Engineering, Information Technology, e-Business, and Applications. Institute (SEITI), Central Michigan University. Foz do Iguazú. Brazil.

Perichinsky, G et al. (2003). Taxonomic Evidence Applying Algorithms of Intelligent Data Mining. Asteroids families. Workshop de Investigadores en Ciencias de la Computación. Red de Universidades Nacionales con Carreras de Informática. Universidad Nacional del Centro. Tandil. Buenos Aires.

Raghavan, V. V. y Birchard, K. (1978). A clustering strategy based on a formalism of the reproductive process in natural systems. Proceedings of the 2nd International Conference on Research and Development in Information Retrieval. 10-22.

Sneath, P.H. and Sokal, R.R. (1973). Numerical Taxonomy, Freeman, San Francisco.