Agregación de métricas de Minería de Datos usando funciones de Lógica Continua

Aristides Dasso*, Ana Funes*
***Universidad Nacional de San Luis**
Ejército de los Andes 950
San Luis
Argentina
{arisdas, afunes}@unsl.edu.ar

Resumen

En la línea de investigación aquí presentada, nos ocupamos de la propuesta y aplicación de una Lógica Continua [Duj08] para la evaluación y comparación de técnicas de Minería de Datos.

Palabras clave: *Lógica Continua. Minería de Datos. Técnicas de evaluación. Técnicas de Minería de Datos. Clasificadores. Logic Score of Preference. LSP*.

Contexto

Este trabajo de investigación se encuentra enmarcado dentro del Proyecto de Incentivos código 22/F822: "Ingeniería de Software: Conceptos, Métodos y Herramientas en un Contexto de Ingeniería de Software en Evolución", de la Universidad Nacional de San Luis, en la línea "Métodos Formales y Prototipos Evolutivos", del mismo. Dentro del contexto de desarrollo de métodos y herramientas, esta investigación tiene como objetivo el concretar la construcción de una herramienta de software que sirva para la evaluación de técnicas de Minería de Datos empleando un modelo de agregación de métricas que se basa en el uso de operadores de una Lógica Continua.

Introducción

En el Aprendizaje Automático o de Máquina, que se encuentra en la base de la Minería de Datos, es necesario evaluar la calidad de los modelos y hacerlo de la manera más precisa que sea posible. La etapa de evaluación de los modelos aprendidos es fundamental para poder garantizar la aplicación de los mismos. Sin embargo, establecer medidas justas y exhaustivas no es tarea sencilla.

Por otro lado, en trabajos previos hemos desarrollado modelos para evaluación de distintas técnicas, herramientas, sistemas, ya sea en el área de recursos humanos [DDF00], herramientas de software [DFP04]; [DPS03]; [FDD00]; [FDPS05] o servicios de e-gov [DDF07]; [CDF09]; [DF11], entre otros. Dichos modelos han sido creados aplicando el método LSP (Logic Score of Preferences) el cual hace uso operadores de una Lógica Continua ([Duj08], [Duj07], [Duj96], [Duj97], [DB97]).

En el presente trabajo apuntamos a integrar (agregar) en un solo valor, haciendo uso de dicha lógica, los distintos resultados que pueden obtenerse al usar métricas diversas cuando se evalúan técnicas de Minería de Datos; particularmente, en este caso, nos atenemos a las técnicas de evaluación de clasificadores.

Para los clasificadores existen diversas métricas para su evaluación, sin embargo no es aconsejable emplear una sola de ellas ya que es común que un método de clasificación presente buenos resultados en una métrica y malos en otra. Por otro lado, realizar comparaciones de los resultados obtenidos con distintas métricas puede resultar confuso al

momento de analizarlas y compararlas.

Es por eso que creemos conveniente obtener un único valor que integre los distintos resultados de las métricas empleadas.

Asimismo, consideramos importante que el método de integración empleado sea comprensible para el evaluador y que, además, le resulte fácil poder adaptarlo a distintas situaciones y técnicas de Minería de Datos.

Algunas de las métricas más empleadas para la evaluación de clasificadores son las técnicas de evaluación de hipótesis basadas en precisión como, por ejemplo, el porcentaje de Error de Muestra (o inversamente el porcentaje de Acierto (Accuracy)), o el Alcance, la Precisión o la Especificidad (Recall, Precision, Specificity) o técnicas más elaboradas como el Área Bajo la Curva (ROC). Estas métricas a su vez pueden ser calculadas empleando ya sea un mismo conjunto de datos tanto para el entrenamiento como para la validación o con conjuntos separados. A su vez puede aplicarse (o no) Validación Cruzada para evitar sesgo en las particiones de los datos. La Validación Cruzada puede aplicarse con o sin solapamiento de los subconjuntos. Otras técnicas de evaluación incluyen la combinación de hipótesis, tales el como el Bootsting, Bagging o Ramdomization.

Resultados y Objetivos

Cabe hacer notar que cada una de las métricas comúnmente usadas en Data Mining provee una perspectiva diferente en el espacio de performance del clasificador, brindando resultados diferentes para un mismo modelo aprendido. Es por eso que creemos conveniente contar con un resultado único proveniente de un modelo comprensible que combine las múltiples métricas pero que a la vez le permita al usuario calibrar dicho modelo de acuerdo a sus necesidades.

Nuestro objetivo es, entonces, integrar los datos obtenidos a partir de diversas métricas de evaluación de clasificadores. Para ello empleamos, como ya se dijo más arriba, una Lógica Continua, específicamente la propuesta en el método Logic Score of Preferences (LSP), cuyo proceso se encuentra resumido en la Figura 1. En ella mostramos una visión global del método LSP con sus correspondientes partes.

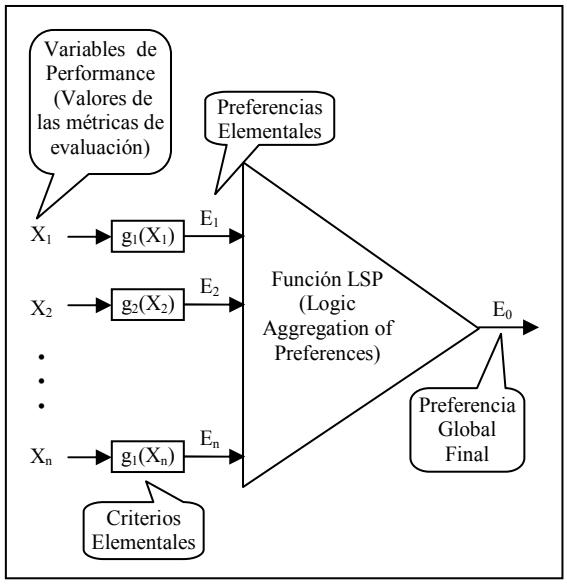

Figura 1. El proceso de evaluación de LSP.

Este método nos permite construir modelos de evaluación que consisten de *Estructuras de Agregación*, las que se basan en el uso combinado de las distintas funciones GCD (Generalized Conjunction Disjunction) del método LSP. Los datos obtenidos por las distintas métricas utilizadas, las cuales de acuerdo a LSP serían las *Variables de Performance* del sistema, son mapeados por medio de *Criterios Elementales* en *Preferencias Elementales*. Son estas preferencias elementales las que van siendo agregadas por medio de funciones GCD en estructuras de agregación y que nos permiten obtener un único valor final.

Los criterios elementales son funciones que transforman un valor real proveniente de una variable de performance en un valor perteneciente al intervalo [0,100]. Estos valores resultantes son llamados preferencias elementales y representan el grado de cumplimiento con un requerimiento del sistema bajo evaluación.

Estos criterios elementales nos permiten establecer, por ejemplo si así se desea, valores mínimos para una variable de performance.

Ilustraremos esto por medio de un ejemplo simple. Supongamos que tenemos un clasificador que debe asignar a cada instancia de una muestra una de dos clases (Positivo / Negativo). Esto nos da la matriz de confusión de la Tabla 1, donde VP: Verdaderos Positivos; FP: Falsos Positivos; FN: Falsos Negativos; VN: Verdaderos Negativos.

Tabla 1. Matriz de confusión

		Predicción	
		Positivos	Negativos
Real	Positivos	VP	FN
	Negativos	FP	VN

Por otro lado, si tenemos las siguientes métricas: Precisión (P): VP / (VP + FP); Alcance/Sensibilidad (AS): VP / (VP + FN); Especificidad (E): VN / (VN + FP); Acierto (A): (VP + VN) / (VP + VN + FP + FN) [LU04], entonces podríamos hacer una estructura de agregación como la que se muestra en la Figura 2.

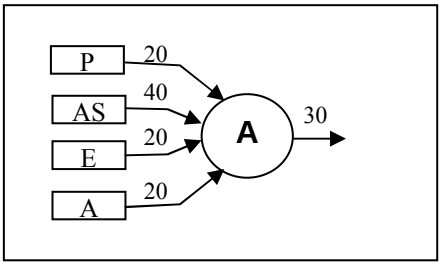

Figura 2. La función **A** (media aritmética) que devuelve la media de los valores de P, AS, E, A ponderada (pesada).

Nótese que en dicha estructura AS (Alcance/Sensibilidad) tiene mayor peso que las otras métricas, esto ha sido hecho nada más que a los efectos de ejemplificar las posibilidades del LSP.

Líneas de Investigación y Desarrollo

Este trabajo es llevado a cabo dentro de la línea de "Métodos Formales y Prototipos Evolutivos" del proyecto de incentivos de la Universidad Nacional de San Luis, código 22/F822: "Ingeniería de Software: Conceptos, Métodos y Herramientas en un Contexto de Ingeniería de Software en Evolución". El mismo se encuentra íntimamente relacionado con trabajos previos en el área del desarrollo de modelos de evaluación. En esta misma línea de investigación, hemos ya aplicado el método LSP para otro tipo de sistemas en trabajos tales como [DF11], [CDF09], [Cas10], [DDF00], [DDF07], [DFP04], [DFPS01], [DPS03], [FDD00], [FDPS05], [MDU00]. El desarrollo y calibración de modelos para la evaluación de clasificadores es uno de los proyectos inmediatos en esta línea, al mismo tiempo que se continúa trabajando en la construcción de modelos de evaluación de calidad de software como otro de los proyectos en esta línea.

A partir de estos desarrollos se analiza la posibilidad de nuevas tesis de grado y posgrado.

Formación de Recursos Humanos

La evaluación de sistemas, métodos y herramientas es una de las áreas en la cual venimos trabajando desde hace varios años y que ha producido varias publicaciones como ya se ha mencionado más arriba. Este trabajo continuo nos ha conducido, entre otros, a la evaluación de sitios de gobierno electrónico, lo que ha dado como resultado tesis de maestría y de licenciatura, a la vez que otras se encuentran en preparación.

Los aspectos propios del trabajo aquí presentado son ambiciosos y se espera que las distintas tareas a desarrollar sirvan para la realización de tesis de posgrado así como de grado.

Referencias

[Cas10] Castro, Marcelo; "Análisis de las propiedades y atributos propios de sitios de gobierno electrónico", Tesis para la Maestría en Ingeniería del Software. Departamento de Informática, Universidad Nacional de San Luis, 2010.

[CDF09] M. Castro, A. Dasso, A. Funes. "Modelo de Evaluación para Sitios de Gobierno Electrónico". 38 JAIIO/SIE 2009, Simposio de Informática en el Estado 2009, Mar del Plata, Argentina, August 26-28, 2009.

[DB97] J. J. Dujmovic and A. Bayucan, "Evaluation and Comparison of Windowed environments", Proceedings of the IASTED Interna Conference Software Engineering (SE'97), pp 102-105, 1997.

[DDF00] N. Debnath, A. Dasso, A. Funes, G. Montejano, D. Riesco, R. Uzal, "The LSP Method Applied to Human Resources Evaluation and Selection", Journal of Computer Science and Information Management, Publication of the Association of Management/International Association of Management, Volume 3, Number 2, 2000, ISBN 1525-4372, pp.1-12.

[DDF07] Narayan Debnath, Aristides Dasso, Ana Funes, Roberto Uzal, José Paganini. "E-government Services Offerings Evaluation Using Continuous Logic". 2007 ACS/IEEE International Conference on Computer Systems and Applications, AICCSA '2007, Amman, Jordan. Sponsored by IEEE Computer Society, Arab Computer Society, and Philadelphia University, Jordan. May 13-16, 2007

[DF11] Aristides Dasso, Ana Funes. "A Model for E-voting Systems Evaluation". 40 JAIIO/SIE 2011, August 29 to September 2, 2011. Córdoba, Argentina.

[DFP04] A. Dasso, A. Funes, M. Peralta, C. Salgado, "User Oriented Evaluation Models for DBMSs", 33 Jaiio (ASIS 04), Córdoba, Argentina, 20-24 de Septiembre, 2004.

[DFPS01] A. Dasso, A. Funes, M. Peralta, C. Salgado, "Una Herramienta para la Evaluación de Sistemas", Workshop de Investigadores en Ciencias de la Computación, WICC 2001, Universidad Nacional de San Luis, San Luis, Argentina, May 2001.

[DPS03] N. Debnath, M. Peralta, C. Salgado, A. Funes, A. Dasso, D. Riesco, G. Montejano, R. Uzal, "Web Programming Language Evaluation using LSP", CAINE03 Proceedings, Las Vegas, USA, 11-13 de Noviembre, 2003. ISBN: 1-880843-49-8, pp 302-305.

[Duj 08] Dujmovic, J.J., "Characteristic forms of generalized conjunction/disjunction"; En Fuzzy Systems, 2008 (FUZZ-IEEE 2008). (IEEE World Congress on Computational Intelligence). 1-6 June 2008, pp. 1075 – 1080, ISSN: 1098-7584, E-ISBN: 978-1-4244-1819-0, Print ISBN: 978-1-4244-1818-3.

[Duj07] Jozo J. Dujmovic, "Continuous Preference Logic for System Evaluation", IEEE Transactions on Fuzzy Systems, Vol. 15, Nº 6, December 2007.

[Duj96] J. J. Dujmovic, "A Method for Evaluation and Selection of Complex Hardware and Software Systems", The 22nd International Conference for the Resource Management and Performance Evaluation of Enterprise Computing Systems. CMG96 Proceedings, vol. 1, pp.368-378, 1996.

[Duj97] J. J. Dujmovic, "Quantitative Evaluation of Software", Proceedings of the IASTED International Conference on Software Engineering, edited by M.H. Hamza, pp. 3-7, IASTED/Acta Press, 1997.

[FDD00] A. Funes, A. Dasso, J. Dujmovic, G. Montejano, D. Riesco, R. Uzal, "Web Browsers Performance Analysis using LSP Method", Proceedings de la International Conference on Software Engineering Applied to Networking & Parallel/Distributed Computing (SNPD'00), Mayo, 2000, Reims, Francia. ISBN: 0-9700776-0-2, pp 551-558.

[FDPS05] Ana Funes, Aristides Dasso, Carlos Salgado, Mario Peralta, "UML Tool Evaluation Requirements". Argentine Symposium on Information Systems ASIS 2005. Rosario, Argentina. September 29-30, 2005.

[LU04] Z. Lu et al., Predicting Subcellular Localization of Proteins using Machine-Learned Classifiers, Bioinformatics, Volume 20, Issue 4, March 2004, pp. 547 - 556.

[MDU00] G. Montejano, J.J. Dujmovic, R. Uzal, D. Riesco, A. Dasso, A. Funes, "A Prototype Tool for Decision Support based in the LSP Method", Proceedings de IASTED, Las Vegas, Nevada, USA, 6-9 de Noviembre, 2000. ISBN: 0-88986-306-7, pp 1-4.