Integrando Enfoques de Medición y Evaluación con Minería de Datos y Procesamiento de Flujos

Mario Diván[1 2], **Luis Olsina**[2]

[1] Facultad de Ciencias Económicas y Jurídicas,
[1 2] GIDIS_Web, Facultad de Ingeniería
Universidad Nacional de La Pampa
[mjdivan, olsinal]@ing.unlpam.edu.ar / mjdivan@eco.unlpam.edu.ar

Resumen. *Este línea de trabajo aborda la problemática de los modelos de clasificación aplicados a flujos continuos de datos, variantes en el tiempo y semi-estructurados (según se define en [1]), usando el marco conceptual de medición y evaluación C-INCAMI (Context - Information Need, Concept model, Attribute, Metric and Indicator [2,3]). Esta investigación integra ambos enfoques, con el fin de generar y soportar un modelo de decisión adaptable al vuelo, que a su vez contribuya al proceso de toma de decisiones en diferentes contextos.*

Palabras clave: *modelos de clasificación, flujos de datos, métricas, indicadores, C-INCAMI.*

1. Introducción

En la actualidad, un nuevo tipo de aplicaciones intensivas en el procesamiento de datos requiere un tratamiento diferenciado del enfoque tradicional de análisis de datos basado en persistencia [4]. Este nuevo tipo de aplicaciones necesita un procesamiento "al vuelo" (on-line) capaz de poder tomar decisiones o ajustar modelos de soporte a la toma de decisiones al momento en que el dato arriba y es procesado, sin disponer de tiempo y/o recursos para un procesamiento secundario en un entorno de persistencia [5].

Bajo el supuesto de que los datos XML arriban a través de varios flujos de datos [6] estructurados a partir del marco formal de medición y evaluación C-INCAMI, se expondrá el foco central del trabajo en base a la generación y ajuste de modelos de clasificación basados en conocimientos preexistentes (datos y metadatos [7]). El objeto de este último punto es brindar un soporte más robusto y consistente al proceso de toma de decisiones sobre los contextos [8] a los que se está tratando de medir y/o evaluar mediante C-INCAMI.

El presente artículo se organiza en tres secciones. La sección 2 aborda los cambios de contextos que afectan al modo y esquema de procesamiento de datos y como éstos repercuten en el proceso de toma de decisiones. La sección tres presenta un modelo de procesamiento de flujos continuos y dentro de sus sub-secciones, se expondrán las funcionalidades asociadas a cada elemento del modelo. Finalmente, en la sección cuatro se esbozan algunas conclusiones y trabajos a futuro.

2. Motivación

La idea de procesar los datos "al vuelo" implica una clara diferenciación con respecto a los sistemas de gestión de bases de datos (SGBD) [9] según el enfoque tradicional. A los efectos de diferenciar el contexto de datos tradicional y el denominado "data streams" (flujos de datos), se debe tener en cuenta, por ejemplo, el problema de la *detección de fraudes* en las empresas de telefonía [10].

Dichas empresas desearían seguramente poder detectar los potenciales fraudes al momento en que se están por producir, o bien localizarlos mientras ocurren, para de este modo poder minimizar las posibles pérdidas asociadas. Esto último motiva a la incorporación de técnicas provenientes del campo de la minería de datos sobre el área de flujos de datos –mining data streams- a los efectos de medir y evaluar el comportamiento típico sobre los flujos de datos asociados a las comunicaciones y poder construir modelos de decisión ajustables "al vuelo" capaces de interpretar y detectar desvíos en las métricas e indicadores definidos. Precisamente, la motivación en usar el marco de medición y evaluación C-INCAMI, es que permite especificar los datos y metadatos de las métricas e indicadores en cuestión, además de las propiedades de contexto.

Por otro lado, si se toma un contexto basado en modelos tradicionales de persistencia de datos, aún haciendo empleo de técnicas de minería de datos, se estarían analizando datos históricos y en caso de que se pudiese localizar un fraude, esto sólo representaría un hecho pasado y en donde el costo asociado al mismo se habrá tornado en una pérdida.

Por lo tanto, la idea de adaptar técnicas de clasificación [11] a flujos de datos semi estructurados enmarcados dentro de C-INCAMI, se asocia a generar y ajustar los modelos de modo que eviten o minimicen dichas pérdidas, sin necesidad de analizar datos históricos sino por el contrario, analizar datos en el mismo momento en que se generan o arriban, pudiendo adaptar los modelos y tomar decisiones más robustas en base al contexto vinculado con los datos.

3. Componentes del Modelo de Procesamiento

Conceptualmente, la idea en términos de procesamiento (ver Fig. 1) es la siguiente. A partir del ingreso de los datos a una función $F^t(d_{si})$ –que tendrá por objeto la suavización de los mismos- se aplicará un modelo actual de clasificación $G(d^t_{si})$ en base al marco de métricas e indicadores definidos junto con el conocimiento previo almacenado, produciendo una decisión en tiempo t, a la cual se denomina D^t. Luego, el modelo actual de clasificación se *ajusta y/o sustituye* en base a los nuevos datos y situación contextual para producir la decisión en tiempo t+1 con el nuevo modelo, permitiendo la comparación de decisiones entre D^t y D^{t+1}. Ambas decisiones permitirán pro-activamente disparar alarmas, notificaciones, ajustes y/o lo que el usuario de C-INCAMI haya definido en términos de indicadores, a los efectos de detectar las desviaciones on-line.

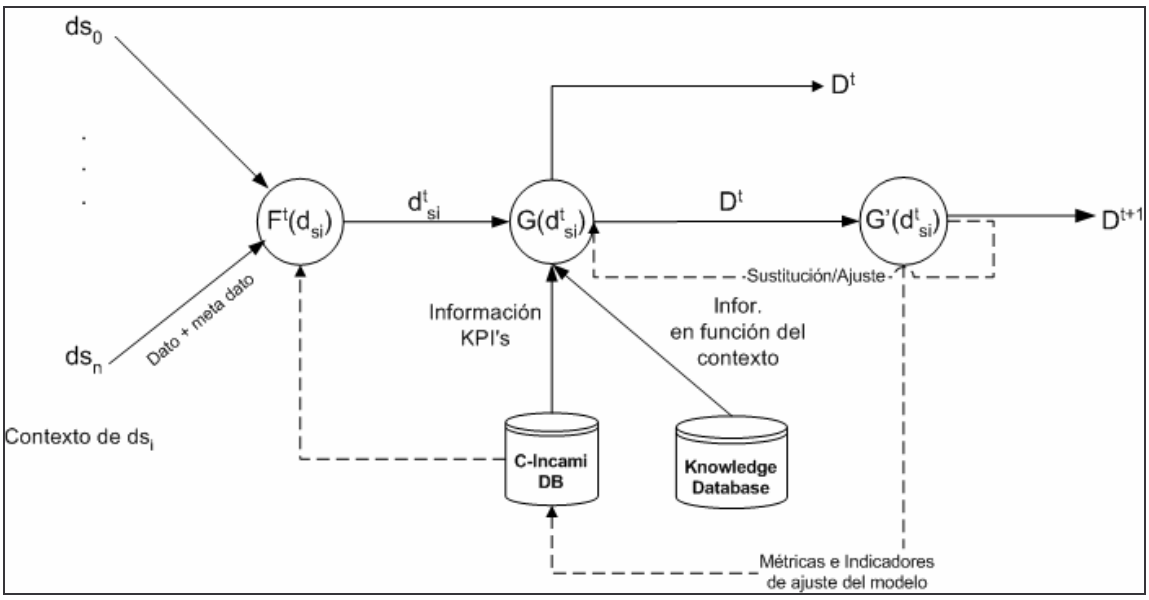

Figura 1. *Esquema del modelo integrado de procesamiento*

Donde:

d_{si}	Flujo de datos (data set) 'i'
$F^t(d_{si})$	Función de transformación o suavizado del d_{si}
d^t_{si}	Flujo de datos 'i' suavizado
$G(d^t_{si})$	Aplicación del modelo de clasificación actual (por ej. árbol)
D^t	Decisión en tiempo 't' obtenida mediante el modelo actual
D^{t+1}	Decisión en tiempo 't+1' obtenida mediante el modelo ajustado
$G'(d^t_{si})$	Sustitución/Ajuste del modelo actual a partir de la generación del nuevo modelo.
KPI	Key Process Indicador (Indicador de proceso clave)
Knowledge DB	Base de datos de conocimiento

3.1 Fuentes de Datos

Se entiende por *fuentes de datos* –data sources- a cualquier origen plausible de generar nuevos datos de un modo continuo e ilimitado. Se entiende por *nuevos datos* a aquellos que no existen previamente en la memoria de procesamiento en el momento exacto en que arriban al procesador; pero en caso de que existiesen, su arribo aporta algún elemento adicional que permite incorporar un nuevo punto de vista al análisis automático de datos. Las fuentes de datos para el contexto de procesamiento en continuo se asumen como *ilimitadas*, que a su vez pueden aprovisionar los datos de a *ráfagas*. Esta última característica implica que la fuente de datos puede exponer comportamiento anómalos, o tal vez por tareas programadas de mantenimiento, que represente momentáneamente la falta de datos en las etapas posteriores de aprovisionamiento.

Las datos se procesan según el concepto de *ventanas* –windows- [12] surgido a partir de la idea de data streams. El procesamiento por ventanas puede ser en dos modos: el primero implica una *ventana en función temporal* que procesará tantos registros como arriben en un período "p"; mientras que el segundo implica una *ventana en función de hitos*; esto es, aquella que procesará tantos datos como existieran dentro de las ventanas definidas por dichos hitos. De este supuesto se desprende como corolario que *los flujos de datos son parcelizables*.

Por otra parte, los datos no arriban conceptualmente en forma unitaria y vendrán acompañados, adicionalmente, de metadatos que permitirán discernir entre su semántica [2]. Al igual que ocurre en el modelo relacional de datos, es posible que no se disponga de valor para un atributo determinado en una medición dada. Además de la ausencia de valor para un atributo en la medición, debe considerarse que es factible que las líneas de comunicación de datos se dañen y se interrumpa el flujo de datos y metadatos total o parcialmente al procesador. La información asociada a los metadatos se estructurará en base a C-INCAMI sobre un esquema XML que permita automatizar su identificación y procesamiento automático.

Por último, y no por ello menos importante, se debe tener en cuenta dentro del flujo de datos asociado a la medición de un atributo, que existen mediciones en las que pueden obtenerse varios valores con sus respectivas probabilidades asociadas, para un momento y contexto dados.

3.2 Función de Suavizado

La función $F^t(d_{si})$ tiene asociada dos responsabilidades. En primer lugar, debe dar coherencia al orden de procesamiento y, en segundo lugar, debe suavizar [13, 14] cada uno de los flujos de datos acorde al procesamiento por ventanas indicado anteriormente.

La coherencia en el orden de procesamiento estará en función de los metadatos que arriben en forma conjunta con los datos, mientras que la suavización de los datos tiene por objeto identificar y resolver inconvenientes propios del dato original tal como el ruido, outliers, y ausencia de datos.

3.3 Función de Decisión

La función de decisión "t", tomará el orden de procesamiento y los datos suavizados para aplicar el modelo de clasificación actual [15] y así obtener la decisión en el tiempo "t". Esta etapa comunica los datos en su estado original, el orden de procesamiento y la decisión en tiempo "t" a la etapa

asociada a la generación del nuevo modelo de clasificación. La etapa de generación del nuevo modelo de clasificación permitirá retroalimentar el comportamiento de la función de decisión "t", brindándole información sobre el nuevo modelo de clasificación generado junto con la decisión "t+1".

3.4 Función de Sustitución/Ajuste del Modelo de Clasificación

La función recibirá la decisión en tiempo "t", el orden de procesamiento y los datos en su estatus original [16] para la aplicación de diferentes técnicas, a los efectos de ajustar y/o sustituir el modelo de clasificación actual.

El contexto de los datos como su naturaleza, influyen empíricamente en el modelo de clasificación a adoptar y/o ajustar. Cuando se habla de su influencia empírica, se hace referencia a la utilización de la base de datos de conocimiento con los modelos históricos asociados a cada uno de sus contextos, como así también a los datos de las mediciones, dentro de C-INCAMI DB para incorporarse dentro del proceso de ajuste/sustitución.

Existen numerosas técnicas para obtener nuevos modelos de clasificación que van desde los tradicionales árboles hasta complejas redes neuronales [17]. Cabe resaltar que no es simple indicar qué técnica es mejor que otra, debido a que las mismas son sensiblemente dependientes del contexto de aplicación, así como de la naturaleza de los datos. Esto último representa un desafío no menor a los efectos de automatizar el proceso de generación del modelo de clasificación y de la decisión de *cuál de los modelos obtenidos aplicar* a los efectos de sustituir/ajustar el modelo de la etapa "t" [18]. Para esta investigación, se ha decidido acotar la función de clasificación sólo a técnicas y métodos asociados a árboles a los efectos de no extender el límite de estudio.

Una vez obtenido el nuevo modelo de clasificación, éste se comunica al proceso de la función de decisión en tiempo "t" junto con la decisión en tiempo "t+1", para actualizar el modelo y así ser empleado en el procesamiento de las ulteriores ventanas. Se entiende que la decisión en tiempo "t+1" se obtiene mediante la aplicación del nuevo modelo de clasificación. Finalmente, se compara la semántica de las decisiones producidas en tiempo "t" y "t+1" y aquella que mejor se adecue globalmente a las condiciones actuales del contexto, será la adoptada.

Esto último presenta interrogantes no tan simples de responder, por ejemplo: ¿Cómo saber cuál decisión es la que mejor se adecua al contexto actual? ¿Qué parámetros regirán el comportamiento del contexto? ¿Todos los parámetros tendrán el mismo peso o existirá algún mecanismo de retroalimentación de los pesos de decisión? Posibles soluciones a estos desafíos serán ampliados en trabajos futuros.

4. Conclusiones y Trabajo Futuro

El presente trabajo ha presentado la propuesta sobre el modelo de procesamiento que permite enfocar el abastecimiento de datos y metadatos como un continuo procesamiento a través de esquemas de parcelización, los cuales permiten construir y/o adecuar modelos de clasificación online para dar soporte al proceso de toma de decisiones en cada contexto. El modelo se sustenta en la especificación de datos y metadatos del marco de medición y evaluación C-INCAMI, como así también en técnicas de clasificación y procesamiento de flujos de datos. Actualmente, los esfuerzos están vinculados a la tarea de definición y consenso de la estructura XML que se adecue al marco, y que permita la serialización/deserialización de datos/metadatos para el procesamiento. Como producto final de la primera etapa, se contará con un prototipo de software que permita generar datos de pruebas basados en metadatos C-INCAMI.

Como trabajo a futuro, se abordarán algoritmos y técnicas de minería de datos con el objeto de la suavización de la serie de datos y la determinación del orden de procesamiento online; como así también la generación y ajuste de modelos de clasificación basado en árboles, la comparación de dichos modelos y el análisis de viabilidad de aplicar uno u otro en base a la situación contextual.

Referencias

1. Chaudhry N., Shaw K., and Abdelguerfi M. (2005). Stream Data Management. Springer. pp. 1-11.

2. Olsina L, Papa F., Molina H. (2007) How to Measure and Evaluate Web Applications in a Consistent Way. Chapter 13 in Springer Book, Human-Computer Interaction Series, titled *Web Engineering: Modelling and Implementing Web Applications*; Rossi, Pastor, Schwabe, & Olsina (Eds.), pp. 385–420.

3. Olsina L, and Molina H. (2007). Towards the Support of Contextual Information to a Measurement and Evaluation Framework. In proc. of 6th Int'l Conference on the Quality Information and Communications Technology (QUATIC07). IEEE CS Press, Lisbon, Portugal. pp. 154–163.

4. Babcock B., Babu S., Datar M., Motwani R., and Widom J. (2002) Models and Issues in Data Stream Systems. In proc. of 21st ACM Symposium of Principles of Databse Systems (PODS 2002). Madison, USA.

5. Fan W., Huang Y., Wang H. and Yu P. (2004). Active Mining of Data Streams. In proc. of Int'l Conference on Data Mining (SIAM2004). Florida, USA.

6. Bose S. and Fegaras L. (2004). Data Stream Management for Historical XML Data . In proc. of Int'l Conference on Management of Data (ACM SIGMOD2004). Paris, France.

7. Medhat Gaber M., Zaslavsky A. and Krishnaswamy S. (2005). Mining Data Streams: A Review. ACM SIGMOD Record, Vol. 34: 2, pp. 18-26, ISSN 0163-5808.

8. Singh S., Vajirkar P. and Lee Y. (2003). Context-Based Data Mining Using Ontologies. Chapter 17th in Springer Book, LNCS titled *Conceptual Modeling – ER 2003*; Song, Liddle, Ling & Scheuermann (Eds.), pp. 405-418.

9. Golab L & Oszu T. (2003). Issues in Data Stream Management. ACM SIGMOD Record, Vol. 34: 2, pp.5-14, ISSN 0163-5808.

10. Abidogun O. (2005). Data Mining, Fraud Detection and Mobile Telecommunications: Call Pattern Analysis with Unsupervised Neural Networks. Master thesis. University of the Western Cape.

11. The Y., Zaitun A., and Lee S. (2001). Data Mining Using Classification Techniques in Query Processing Strategies. ACS/IEEE Int'l Conference on Computer Systems and Applications (AICCSA'01).

12. Tao Y., and Papadias D. (2006). Maintaining Sliding Window Skylines on Data Streams, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18: 3, pp. 377-391.

13. Pérez López C. (2005). Métodos Estadísticos Avanzados con SPSS. Thomson.

14. Johnson D. (2000). Métodos Multivariados Aplicados al Análisis de Datos. Thomson.

15. Aggarwal C., Han J., Wang J. and Yu, P. S. (2004). On Demand Classification of Data Streams, Proc. of Int'l Conf. on Knowledge Discovery and Data Mining (KDD'04), Seattle, WA.

16. Ben-David S., Gehrke J., and Kifer D. (2004). Detecting Change in Data Streams. Proc. of VLDB04.

17. Dong G., Han J., Lakshmanan L.V.S., Pei J., Wang H. and Yu P.S. (2003). Online mining of changes from data streams: Research problems and preliminary results. In Proc. of the Workshop on Management and Processing of Data Streams. In cooperation with the Int'l Conference on Management of Data (ACM-SIGMOD'03), San Diego, CA.

18. Gama J., Medas P., and Rodríguez P. (2005). Learning Decision Trees from Dynamic Data Streams, ACM Symposium on Applied Computing - SAC05.