IX Jornadas de Ciencia y Tecnología
de Facultades de Ingeniería del NOA
Santiago del Estero, 3 y 4 de Octubre de 2013

Utilización de una arquitectura para data mining stream en el análisis de datos correspondiente a precipitaciones

Nélida R. Cáceres[1] & Viviana E. Quincoces[1]

(1) Facultad de Ingeniería, Universidad Nacional de Jujuy
{nrcaceres, vquincoces}@fi.unju.edu.ar

RESUMEN: En la actualidad surge el desafío para la Minería de Datos (data mining) de trabajar no sólo con datos históricos, sino también con flujos de datos (data stream) provenientes de dispositivos utilizados cada vez con mayor frecuencia y que adquieren información del contexto: dispositivos móviles, de posicionamiento global, sensores, entre otros. Esta tecnología debe adecuarse a los cambios y preveer los métodos para capturar los datos del entorno y analizarlos, de forma tal que no obstruya las operaciones cotidianas de una organización. Existen dos arquitecturas para realizar el tratamiento de este tipo de información: distribuida y centralizada. En este trabajo se aplica una técnica de data mining a datos de precipitaciones ocurridas y relevadas en estaciones meteorológicas ubicadas en la provincia de Jujuy (Rca. Argentina) a ambas arquitecturas. Los resultados de la experiencia manifiestan mayor concordancia en el caso de aplicar la arquitectura distribuida.

1. INTRODUCCIÓN

La arquitectura para minería de data stream distribuida presentada por Parthasarathy *et al*. (2007), propone que los data stream que ingresan de forma continua y de distintos nodos reciban un tratamiento adecuado que permita su análisis sin obstaculizar las actividades diarias de una organización. Es un reto para la Minería de Datos (MD) el aplicar técnicas en tiempo real a los datos que van ingresando de forma continua y obtener patrones representativos de ellos.

En este trabajo aplican técnicas de MD de forma distribuida y centralizada, lo que va a permitir realizar comparaciones de los resultados obtenidos y demostrar si el aplicar data mining de forma descentralizada proporciona un beneficio al almacenamiento de datos, a la utilización de los recursos y a la fiabilidad de los patrones obtenidos.

Para realizar MD, se consideraron datos de precipitaciones provenientes de tres estaciones meteorológicas ubicadas en la región del valle de la Pcia. de Jujuy, las mediciones que se realizaron diariamente fueron totalizadas al finalizar el día, y para este trabajo solamente se consideraron los días en los que hubo ocurrencia de precipitaciones. Estos datos constituyeron tres datawarehouse correspondientes con cada una de las estaciones meteorológicas.

La validez de esta arquitectura será probada, en primer término con datos históricos provenientes de fuentes de datos continuas, para determinar las ventajas y desventajas del modelo presentado, quedando para posteriores trabajos realizar MD a los data stream que van ingresando a la organización. En el apartado 2 se describe el marco teórico, en el apartado 3 se describen los datos que se utilizarán para probar la arquitectura mediante una técnica de MD, en el apartado 4 y 5 se presentan respectivamente, las conclusiones y las referencias.

2. MARCO TEÓRICO

2.1 Data Mining (MD)

Witten & Frank (2005), definen MD como el proceso de descubrir patrones en los datos. El proceso debe ser automático o semiautomático (más frecuente). Los patrones descubiertos deben ser significativos ya que dan lugar a ciertas ventajas (por lo general una ventaja económica). Los datos están siempre presentes en cantidades considerables.

Para Hernández Orallo *et al*. (2004), los datos en MD pasan de ser un producto (el resultado histórico de los sistemas de información) a constituir una materia prima que hay que explotar para obtener el verdadero producto elaborado, que es el conocimiento, especialmente valioso para la ayuda en la toma de decisiones en el ámbito en el que se han recopilado o extraído los datos.

Basándose en estos conceptos se define MD como una tecnología que consiste en extraer patrones o modelos, mediante distintas técnicas, que son aplicadas dependiendo del conocimiento que se pretende descubrir, en grandes cantidades de datos históricos y estáticos almacenados en repositorios de datos, generalmente denominados datawarehouse (DW). Posteriormente estos patrones son analizados para extraer conocimiento que sirve de apoyo a la toma de decisiones estratégicas en el ámbito del cual son originarios. El DW contiene datos históricos provenientes de fuentes internas o externas de datos, este conjunto de datos se encuentra adecuado e integrado para que se puedan aplicar las distintas técnicas de MD (Fig. 1). La constitución de este tipo de almacenamiento permite que las transacciones normales de una organización no se vean perturbadas por el análisis de estos datos con la mencionada tecnología.

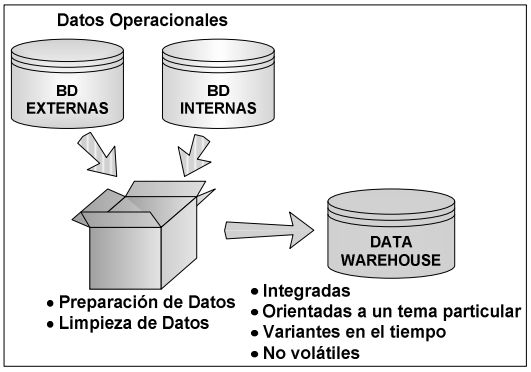

Figura 1. Integración de un Datawarehouse

2.2 Data Stream (flujo de datos)

Tatbul *et al*. (2003) define que un Data Stream (DS) o flujo de datos es una fuente continua y rápida de datos de una variedad de fuentes como sensores, dispositivos de geo-posicionamiento, o programas de computadora. Las nuevas aplicaciones requieren monitorización en tiempo real de flujos de datos. Los sistemas para gestionar estos flujos de datos han surgido para permitir un procesamiento eficiente para atender las necesidades de dichas aplicaciones. Uno de los principales retos en la gestión del flujo de datos es apoyar el procesamiento de los datos en tiempo real con las limitaciones de los recursos del sistema como la CPU, la memoria o ancho de banda. Con un gran flujo de datos y consultas continuas, el sistema experimenta escasez de recursos computacionales produciendo demora en el procesamiento de datos. Para evitar este tipo de problemas el sistema debe controlar la carga de datos, restringiéndola cuando la demanda de recursos es superior a su capacidad. Pero este desprendimiento de carga sacrifica la exactitud de las respuestas del sistema ante una consulta lo que produce degradación en la precisión.

Para Guha *et al.* (2003), los problemas de la MD en la actualidad se refieren a un nuevo tipo de datos, llamados flujos de datos o data stream, que constituye una secuencia ordenada de puntos que se pueden leer solamente una vez o un pequeño número de veces; estos datos provienen de sensores, de los clics en la web, de multimedia, etc. Estos conjuntos de datos son generalmente demasiado grandes para permanecer en memoria, a veces, incluso, exceden la capacidad de los discos. Por consiguiente, la entrada de datos para los algoritmos de DS debe realizarse en línea y no almacenar toda la información para futuros análisis. Estos nuevos modelos dentro del paradigma de data stream deben trabajar con datos en movimiento o compactando la descripción de los datos históricos para utilización futura.

En Aggarwal (2007) se menciona que los avances en la tecnología de hardware han facilitado la posibilidad de recopilar datos de forma continua. Las transacciones sencillas de la vida cotidiana como el uso de una tarjeta de crédito, un teléfono o la navegación por la web permiten el almacenamiento de datos automatizado. Del mismo modo, los avances en la tecnología de la información han dado lugar a gran cantidad de flujos de datos a través de redes IP. Estos grandes volúmenes de datos pueden ser analizados para obtener información interesante y relevante en una amplia variedad de aplicaciones. Con el volumen creciente de los datos, ya no es posible procesar los datos de manera eficiente mediante el uso de múltiples pasadas, más bien, se puede procesar un elemento de datos a lo sumo una vez. Ello supone una presión sobre la aplicación de los algoritmos subyacentes. Por lo tanto, los algoritmos de Stream Data Mining (SDM), deben ser diseñados de manera tal que los algoritmos trabajen mediante una sola pasada de los datos. En muchos casos, no es un componente temporal inherente para el proceso de SDM. Esto es debido a que los datos pueden evolucionar con el tiempo. Este comportamiento de los DS se conoce como localidad temporal. Por lo tanto, una adaptación directa de algoritmos de minería, no puede ser una solución eficaz para la tarea. Algoritmos de SDM deben ser cuidadosamente diseñados con un claro enfoque en la evolución de los datos subyacentes. Otra característica importante de data streams es que a menudo se extrae en una forma distribuida, además los procesadores individuales pueden tener procesamiento y memoria limitada.

2.3 Mining Data Streams (Minería de flujo de datos)

Mining Data Streams se refiere a la extracción de conocimiento representado en modelos o patrones desde flujos de datos (Medhat Gaber *et al*., 2005). Escobar Jeria (2007), define a la Minería Stream como el proceso de extracción de conocimiento de estructuras de registros rápidos y continuos de datos. Esta disciplina debe enfrentarse a los inconvenientes de la memoria limitada debido al rasgo continuo de los elementos entrantes de datos. La aplicación de los algoritmos de MD deben tratar la alta tasa de datos stream generados desde sensores y otras fuentes inalámbricas de datos que crean un desafio real para transferir estas cantidades inmensas de datos a un servidor central para ser analizadas. Algunas de las estrategias para responder a estos desafios incluyen: considerar el muestreo como el proceso que permite seleccionar los DS que serán analizados; considerar la agregación para representar la cantidad de DS de alguna forma estadística que usa elementos agregados, como es el promedio; considerar, para hacer frente a la cantidad de datos entrantes, la clasificación de los DS entrantes en un numero limitado de categorías y reemplazar cada elemento entrante con la categoría más adecuada según una medida especificada. Esto produciría que se conserve la memoria limitada.

En Nasereddin (2009) se expresa que, debido a las características de velocidad tanto offline como online, no hay tiempo suficiente para volver a examinar la BD completa o realizar una nueva exploración con MD cada vez que se produzca una actualización. Además no existe espacio suficiente para almacenar todos los datos resultantes del procesamiento en línea.

2.4 Arquitectura para minería de data stream distribuida

Según (Parthasarathy *et al.*, 2007), aunque el espacio de almacenamiento económico hace que sea posible mantener grandes volúmenes de datos, el acceso y gestión de los datos se convierte en un problema de rendimiento. A menudo se encuentra que un único nodo es incapaz de cubrir los grandes conjuntos de datos, por esto son necesarios técnicas eficientes y adaptables para el acceso, almacenamiento y comunicación de datos (si se distribuyen las fuentes de datos). La MD se hace más complicada en el contexto de bases de datos dinámicas, donde hay un flujo constante de datos. Los cambios en los datos pueden invalidar los patrones existentes o introducir otros nuevos. Volver a ejecutar los algoritmos a partir de cero conduce a grandes gastos computacionales de entrada/salida. Estos dos factores han llevado al desarrollo de algoritmos distribuidos para el análisis de flujo de datos. Muchos sistemas utilizan un modelo centralizado para la minería de data stream. Bajo este modelo de los flujos de datos distribuidos son dirigidos a una ubicación central antes de que se apliquen técnicas de MD (Fig. 2).

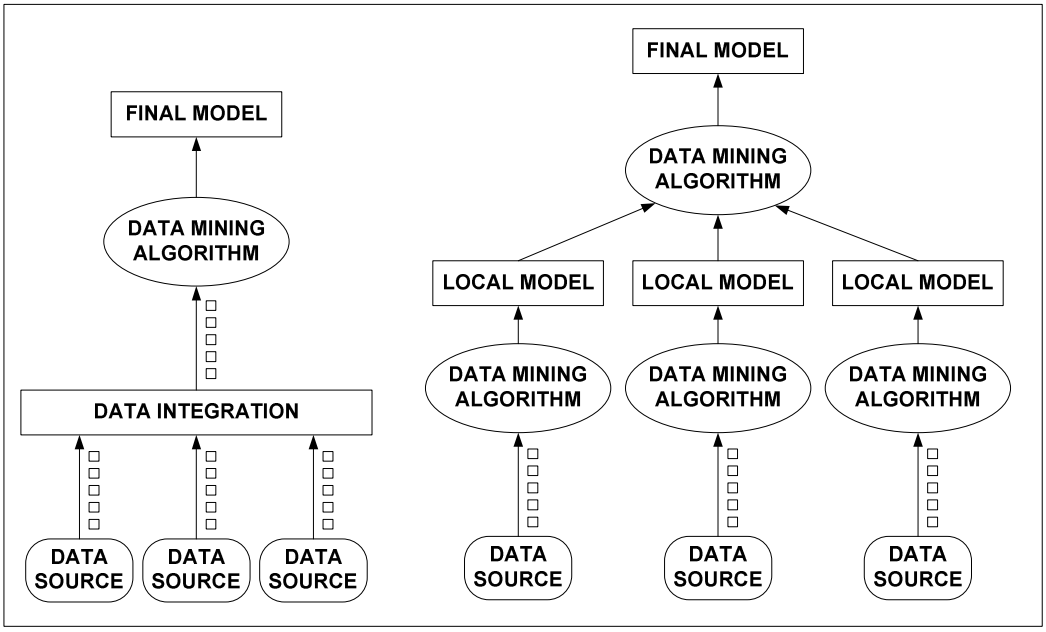

Figura 2. Arquitectura para el proceso centralizado de data stream (izquierda) y arquitectura para el proceso distribuido de data stream (derecha) (Parthasarathy *et al.*, 2007)

La minería centralizada de data stream puede dar como resultado un tiempo de respuesta largo, si bien los recursos informáticos distribuidos pueden estar disponibles, no se utilizan completamente, y el centro de recolección de datos puede dar lugar a tráfico pesado a través de enlaces de comunicación críticos. Si estas vías de comunicación tienen ancho de banda limitado, la red de entrada/salida puede convertirse en un cuello de botella. Por otra parte, en las fuentes de dominios restringidos, tales como redes de sensores, esto puede resultar en un gran consumo de energía debido a la comunicación de datos excesiva. Para reducir los problemas mencionados anteriormente, es que surge un modelo para fuentes distribuidas de datos, recursos computacionales y enlaces de comunicación. En el modelo de minería de data stream distribuida, en lugar de descargar los datos en una ubicación central, los nodos de computación distribuida realizan partes de la computación cerca de los datos, mientras que la comunicación de los modelos locales lo realizan a un sitio central cuando es necesario. Tal arquitectura proporciona varios beneficios, entre los que se pueden mencionar: el uso de nodos de computación distribuida permite la derivación de un mayor grado de paralelismo, reduciendo así el tiempo de respuesta; debido a que solo los modelos locales deben ser comunicados, la comunicación puede potencialmente ser reducida, mejorando la escalabilidad y reduciendo el consumo de energía.

3. UTILIZACIÓN DE LA ARQUITECTURA PARA MINERÍA DE DATA STREAM DISTRIBUIDA EN DATOS DE PRECIPITACIONES OCURRIDAS EN LA PROVINCIA DE JUJUY

Para comparar los resultados obtenidos que surjan de la utilización de la arquitectura para data stream centralizada y distribuida de Parthasarathy *et al.* (2007) se utilizan datos de precipitaciones recabadas diariamente entre setiembre de 1998 y agosto de 2004 en tres estaciones meteorológicas de la Pcia. de Jujuy, ubicados en la región de los Valles de la provincia. Las mediciones fueron realizadas en cada estación con un pluviómetro y la información de precipitaciones fue recopilada en forma continua en cada una de las estaciones, pero al finalizar el día, la información obtenida fue totalizada y almacenada en el datawarehouse. Cabe destacar que solamente figuran para este estudio los días que tienen ocurrencia de precipitaciones. Se consideran importantes los atributos que se describen en la Tabla 1 relacionados con las precipitaciones.
Las estaciones meteorológicas que aportaron la información para este trabajo son las siguientes:

- Augusto Romain de la Facultad de Ciencias Agrarias, ubicada en San Salvador de Jujuy, Dpto. Dr. Manuel Belgrano. (Estación Jujuy Augusto Romain, 2007)
- Los Alisos (Dique), ubicada a 22 km de San Salvador de Jujuy, Dpto. San Antonio, (Estación Los Alisos, 2007)
- El Tipal, ubicada a 27 km de San Salvador de Jujuy, Dpto. El Carmen, (Estación El Tipal, 2007)

Tabla 1. Características de las precipitaciones en la Pcia. de Jujuy

Denominación	*Descripción*	*Tipo de dato*	*Rango de valores*
Día	Dia en que ocurrió la precipitación.	Numérico	1..31
Mes	Mes en que ocurrió la precipitación.	Nominal	Enero, … ,Diciembre
Año	Año en que ocurrió la precipitación.	Numérico	1998, … , 2004
Cantidad de precipitación	Valor diario de la precipitación ocurrida.	Numérico	Mayor o igual que cero (expresado en mm)
Intensidad de la precipitación (FAA, 1986) (Predicciones Meteorólogicas, 2013)	Estimación de la intensidad en base a la cantidad de precipitaciones acaecidas.	Nominal	Niebla (0,05 mm/h) Llovizna(0,25 mm/h) Lluvia ligera (1 mm/h) Lluvia Moderada (4 mm/h) Lluvia Fuerte (15 mm/h) Lluvia Muy Fuerte (40 mm/h) Lluvia Intensa (100 mm/h o más)
Estación del año	Estación del año en que ocurrió la precipitación.	Nominal	Verano, Otoño, Invierno, Primavera

3.1. Data mining stream distribuida y centralizada

En primer lugar se constituyen los DW con datos históricos para cada una de las estaciones meteorológicas, se aplica para este ejemplo, la técnica de de Clasificación Arbol de Decisión J48 del software Weka versión 3.7.2 y se obtienen los patrones que se presentan en las Tablas 2, 3 y 4.
Los patrones resultantes son utilizados para aplicar nuevamente la misma técnica J48 de Weka, para aplicar nuevamente data mining se crea un nuevo atributo denominado "*EstaciónMeteorológica*" del tipo nominal, el cual contiene el nombre de la estación meteorológica a la cual pertenecen los datos. También se crea, por separado otro DW, que se utiliza como repositorio que contiene los datos de las tres estaciones meteorológicas, en este caso también se agrega el atributo que se mencionó anteriormente, se aplica la misma técnica de árbol de decisión, y los resultados obtenidos son comparados en la tabla 5.

Tabla 2. Resultados obtenidos en Estación Meteorológica El Tipal

El Tipal (198 datos)
Estacion_Año = verano | año <= 1998: Lluvia_Moderada (4.0/1.0) | año > 1998: Lluvia_Fuerte (75.0/50.0) Estacion_Año = otoño: Lluvia_Ligera (52.0/23.0) Estacion_Año = invierno: Lluvia_Ligera (14.0/2.0) Estacion_Año = primavera | año <= 1999: Lluvia_Fuerte (28.0/17.0) | año > 1999: Lluvia_Ligera (25.0/15.0)
Kappa statistic….. 0.0736 (*pobre*)

Tabla 3. Resultados obtenidos en Estación Meteorológica Augusto Romain

Augusto Romain (727 datos)
Estacion_Año = verano: Lluvia_Moderada (300.0/199.0) Estacion_Año = otoño | año <= 2000 | | año <= 1999: Lluvia_Moderada (38.0/22.0) | | año > 1999: Lluvia_Ligera (28.0/16.0) | año > 2000: Lluvia_Ligera (118.0/71.0) Estacion_Año = invierno | año <= 1998: Lluvia_Moderada (3.0) | año > 1998: Lluvia_Ligera (66.0/34.0) Estacion_Año = primavera | año <= 2001: Lluvia_Moderada (116.0/71.0) | año > 2001: Lluvia_Ligera (58.0/41.0)
Kappa statistic…. 0.072 (*pobre*)

Tabla 4. Resultados obtenidos en Estación Meteorológica Los Alisos

Los Alisos (577 datos)
Estacion_Año = verano: Lluvia_Moderada (234.0/135.0) Estacion_Año = otoño | año <= 2000 | | año <= 1999: Lluvia_Moderada (40.0/21.0) | | año > 1999: Lluvia_Ligera (27.0/16.0) | año > 2000: Lluvia_Ligera (86.0/41.0) Estacion_Año = invierno: Lluvia_Ligera (45.0/20.0) Estacion_Año = primavera | año <= 1998: Lluvia_Ligera (28.0/17.0) | año > 1998 | | año <= 2001 | | | año <= 1999: Lluvia_Moderada (22.0/11.0) | | | año > 1999: Lluvia_Ligera (46.0/26.0) | | año > 2001: Lluvia_Moderada (49.0/26.0)
Kappa statistic….. 0.0181 (*pobre*)

Tabla 5. Comparación de resultados al utilizar J48 de forma distribuida y centralizada

Datos distribuidos (84 datos)	*Datos centralizados (1502 datos)*
Estacion_Año = verano | EstacionMeteo = SSdeJujuy: Lluvia_Moderada (7.0) | EstacionMeteo = El_Tipal: Lluvia_Fuerte (7.0/1.0) | EstacionMeteo = Los_Alisos: Lluvia_Moderada (7.0) Estacion_Año = otoño | año <= 1999 | | EstacionMeteo = SSdeJujuy: Lluvia_Moderada (2.0) | | EstacionMeteo = El_Tipal: Lluvia_Ligera (2.0) | | EstacionMeteo = Los_Alisos: Lluvia_Moderada (2.0) | año > 1999: Lluvia_Ligera (15.0) Estacion_Año = invierno: Lluvia_Ligera (21.0/1.0) Estacion_Año = primavera | EstacionMeteo = SSdeJujuy | | año <= 2001: Lluvia_Moderada (4.0) | | año > 2001: Lluvia_Ligera (3.0) | EstacionMeteo = El_Tipal | | año <= 1999: Lluvia_Fuerte (2.0) | | año > 1999: Lluvia_Ligera (5.0) | EstacionMeteo = Los_Alisos | | año <= 2000: Lluvia_Ligera (3.0/1.0) | | año > 2000: Lluvia_Moderada (4.0)	Estacion_Año = verano: Lluvia_Moderada (613.0/387.0) Estacion_Año = otoño | año <= 2000 | | año <= 1999: Lluvia_Moderada (112.0/63.0) | | año > 1999: Lluvia_Ligera (55.0/32.0) | año > 2000: Lluvia_Ligera (222.0/117.0) Estacion_Año = invierno: Lluvia_Ligera (128.0/59.0) Estacion_Año = primavera: Lluvia_Moderada (372.0/238.0)
Kappa statistic….. 0.6377 (*buena*)	Kappa statistic….. 0.0689 (*pobre*)

3.2. Interpretación de los resultados obtenidos

El estadístico Kappa, que se destaca en las tablas 2 y 3, hace referencia a la concordancia que presentan los datos analizados. La fuerza de concordancia se puede observar en la siguiente escala (López de Ullibarri Galparsoro *et al.*, 1999) que se muestra en la tabla 4.

Tabla 4. Valoración del índice Kappa

Valor de K	*Fuerza de la concordancia*
Menor a 0.20	Pobre
0.21 - 0.40	Débil
0.41 - 0.60	Moderada
0.61 - 0.80	Buena
0.81 - 1.00	Muy buena

En los datos analizados de forma individual, es decir para cada estación meteorológica, la fuerza de concordancia es "*pobre*", si bien se contaron para cada uno de los casos con una cantidad considerable de datos, por lo que se espera que el índice Kappa mejore al utilizar una mayor cantidad de datos.

Al contar con los patrones resultantes de la aplicación de MD a los datos de cada una de las estaciones meteorológicas, se constituyó un nuevo conjunto de datos representativos para aplicar nuevamente la técnica J48 y obtener el modelo final, produciendo esto que el índice Kappa mejore considerablemente siendo valorado como "*bueno*", a pesar de que no fueron muchos los datos analizados.

En los datos analizados de forma centralizada, es decir que se constituyó un único datawarehouse con los datos de las tres estaciones meteorológicas, la fuerza de concordancia es "*pobre*" a pesar de que al trabajar de esta forma el número de datos aparece incrementado sustancialmente.

Como resultado de utilizar el modelo se resume lo siguiente:

- En el caso de los datos distribuidos se logró un patrón con un mayor grado de concordancia que el conseguido al aplicar la misma técnica de MD a los datos centralizados.
- El árbol que se obtiene con MD distribuida es más detallado y representativo que el que se consiguió con MD centralizada (Fig. 3 y Fig. 4).
- Se comprueba que el índice Kappa mejora cuando se aplica dos veces MD sobre los datos distribuidos que al aplicarlo sólo una vez, es decir, que el tener como referente la arquitectura propuesta mejora el patrón obtenido, contrariamente al supuesto de que el utilizar mayor cantidad de datos mejoraría la concordancia de los datos.
- Se requiere mayor cantidad de recursos al utilizar la arquitectura para extraer patrones, pero los resultados alcanzados son beneficiosos.

Para una mejor comprensión de los resultados obtenidos con la aplicación de esta arquitectura se presenta la misma en la Fig. 5, pero adaptada con los resultados resumidos de todo el proceso de MD que se realizó sobre los datos de precipitaciones.

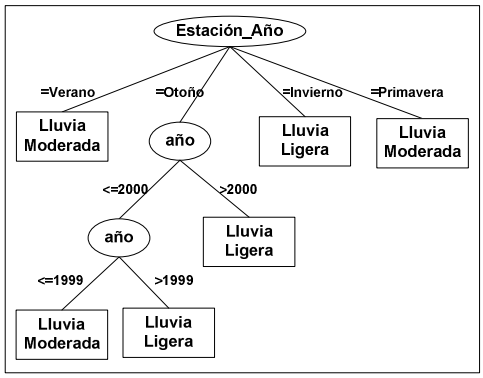

Figura 3. Arbol de decisión al aplicar MD centralizada

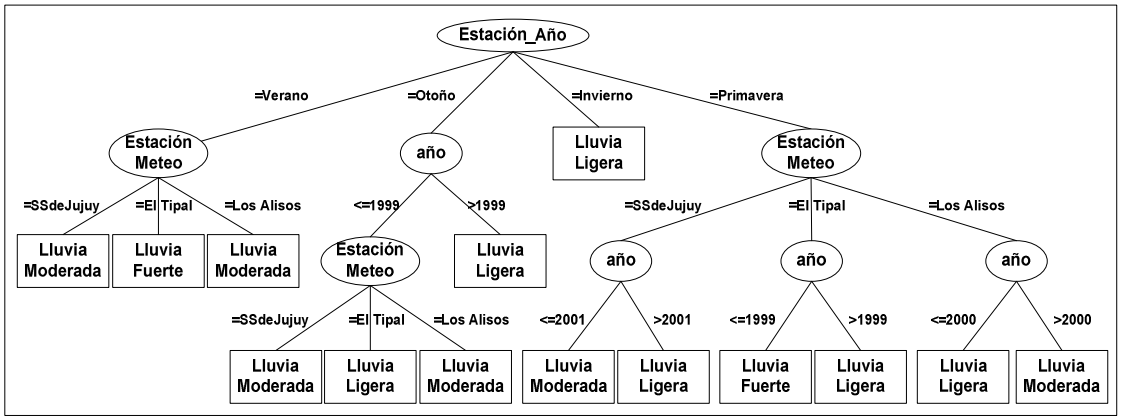

Figura 4. Arbol de decisión al aplicar MD distribuida

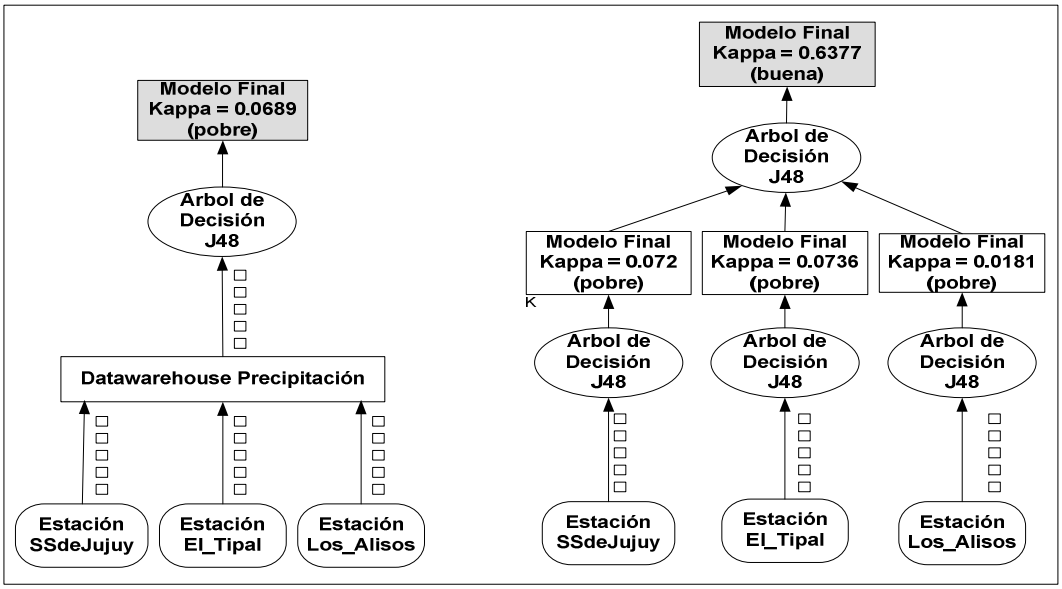

Figura 5. Arquitectura distribuida y centralizada para las precipitaciones en la Pcia. de Jujuy

4. CONCLUSIONES

Para este trabajo se constituyeron tres DW con datos históricos de precipitaciones relevadas en tres estaciones meteorológicas de la Pcia. de Jujuy durante el periodo comprendido entre setiembre/1998 a agosto/2004, se aplicó la técnica J48 de árbol de decisión a los almacenes de datos por separado y luego se conformó un nuevo DW con el conjunto de datos total, de esta forma se realizó MD distribuida y centralizada respectivamente, comprobando que al aplicar técnicas de MD a datos del tipo continuo, los resultados obtenidos al utilizar la arquitectura de minería de data stream distribuida mejora la concordancia de los datos. Si bien el constituir un datawarehouse en forma centralizada y luego aplicar a éste técnicas de MD resulta más práctico y económico (ya que se ahorran recursos), a la hora de obtener resultados, estos son muy resumidos en comparación de los obtenidos de forma distribuida que son más detallados y satisfactorios, con la desventaja de no ahorrar recursos ya que se debe realizar MD dos veces pero que mejoran los resultados. Al realizar la comparación de los patrones obtenidos se puede observar que el aplicar MD de forma distribuida el patrón que se obtiene con menor cantidad de datos tiene una mejor concordancia de los datos (el índice Kappa) que el que se obtiene al aplicar MD centralizada a un gran conjunto de datos, concluyendo, para esto caso, que la precisión del patrón mejora al efectuar dos veces MD sobre un conjunto de datos representativos que al realizar MD sobre gran cantidad de datos. La información utilizafue relevada en estaciones meteorológicas ubicadas en tres localidades de la provincia que pertenecen a la Región del Valle, y es probable que el índice Kappa mejore si se realiza el estudio en estaciones ubicadas dentro de la misma localidad, es decir, al realizar un estudio con datos en una localidad específica. También se puede agregar que esta forma de MD distribuida permite que cada estación meteorológica realice el análisis de sus datos cuando lo crea conveniente, sin esperar que se encuentren centralizados para recién obtener resultados.

Como trabajo futuro se plantea aplicar la arquitectura con datos provenientes desde sensores en tiempo real, trabajar con el datawarehouse ya constituido y efectuar MD de forma más frecuente para optimizar los patrones obtenidos, correspondiendo realizar en este caso MD incremental.

5. REFERENCIAS

Aggarwal, Charu C., *Data Streams: Models and Algorithms,* edited by Springer, NY, USA, 2007.

Escobar Jeria, Victor H., *Minería Web de Uso y Perfiles de usuario: Aplicaciones con lógica difusa,* Editorial de la Universidad de Granada, España, 2007.

Estación El Tipal (El Carmen), División Hidrología, Dirección Provincial. de Recursos Hídricos de la Pcia. de Jujuy, Argentina, acceso 2007.

Estación Jujuy Augusto Romain Convenio UNJU - S.M.N. a cargo de la Cátedra de Agroclimatología de la Facultad de Ciencias Agrarias de la UNJU, Argentina, acceso 2007.

Estación Los Alisos (Dique), División Hidrología, Dirección Provincial. de Recursos Hídricos de la Pcia. de Jujuy, Argentina, acceso 2007.

FAA, Fuerza Aérea Argentina, Boletín N° 30, Publicación impresa en el Servicio Meteorológico Nacional, Argentina, 1986.

Guha, S., Meyerson, A., Mishra, N., Motwani, R., & L. O'Callaghan, *Clustering data streams: Theory and practice*. IEEE Transactions on Knowledge and Data Engineering, 2003.

Hernández Orallo, J., M.J. Ramírez Quintana, & C. Ferri Ramírez, *Introducción a la Minería de Datos*, Departamento de Sistemas Informáticos y Computación, Pearson Educación S.A., Madrid, España, 2004.

López de Ullibarri Galparsoro I. & S. Pita Fernández, *Medidas de concordancia: el índice de Kappa*, *Cad Aten Primaria*, vol 6, 169-171, España, 1999.

Medhat Gaber, M.; Zaslavsky A. & S. Krishnaswamy, *Mining Data Streams: A Review*, Volume 34, Issue 2, Pages 18-26, 2005.

Nasereddin, Hebah H. O, *Stream Data Mining*, Volume 1, Number 4, Pages 183-190, 2009.

Parthasarathy, Srinivasan, Ghoting, Amol & Matthew E. Otey, *Chapter 13: A Survey of distributed mining of data streams*, Department of Computer Science and Engineering The Ohio State University, book Aggarwal, Charu C., *Data Streams: Models and Algorithms,* pages 289-307, edited by Springer, NY, USA, 2007.

Predicciones Meteorológicas - Interpretación, http://www.aemet.es/documentos/es/eltiempo/prediccion/comun/prediccion_interpretacion.pdf acceso Julio 2013.

Tatbul N., Cetintemel U., Zdonik S., Cherniack M. & M. Stonebraker, *Load Shedding on Data Streams Manager,* 2003, http://www.vldb.org /conf/2003/papers/S10P03.pdf, acceso Julio 2012.

Witten I.H. & E. Frank, *Data Mining Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, San Francisco, USA, 2005.