Reducción de datos aplicada a Big Data

Doctorando: Sergio Ramírez Gallego
Directores: Francisco Herrera Triguero
José Manuel Benítez Sánchez

Grupo de Investigación "Soft Computing and Intelligent Information Systems" (SCI^{2}S).
Departamento de Ciencias de la Computación e Inteligencia Artificial,
Universidad de Granada.
sramirez@decsai.ugr.es
Fecha de Inicio: Octubre de 2014

Palabras clave: Reducción de Datos, Grandes Bases de Datos, Sistemas Paralelos, Sistemas Distribuidos, Minería de Datos

1. Introducción

La minería de datos es un campo interdisciplinar con el objetivo general de descubrir conocimiento [**?**]. Las técnicas de minería de datos son sensibles a la calidad de la información sobre la que se pretende extraer conocimiento. Cuanto mayor sea esta calidad, mayor será la de los modelos de toma de decisiones generados. Existe en el proceso de descubrimiento una etapa de pre-procesamiento previa a la minería de datos [**?**], con el objetivo de mejorar la calidad de los datos. Gracias a estas técnicas de pre-procesado los algoritmos de minería de datos puedan obtener modelos que aporten mayor y mejor conocimiento.

Dentro de la etapa del pre-procesamiento destacan los procesos de reducción de datos, donde el objetivo es extraer del conjunto original de datos un nuevo conjunto de datos más pequeño y de mayor calidad para la posterior fase de minería de datos. La reducción de datos se puede llevar a cabo de múltiples formas: selección y extracción de características, discretización, selección y generación de instancias, entre otras.

Actualmente, los sistemas de extracción de conocimiento tienen que soportar ingentes cantidades de datos. El procesamiento de grandes cantidades de información, conocido como Big Data [**?**], no sólo hace mención a grandes volúmenes de datos, sino que también se incluye el proceso de expansión de dichos datos a través de tres dimensiones: volumen, velocidad y variedad.

Una serie de herramientas y plataformas han ido apareciendo con el objetivo de abordan los problemas derivados del uso de grandes cantidades de datos. La plataforma más popular es Hadoop [**?**]; plataforma de código abierto más popular del mercado que sigue el paradigma MapReduce [**?**]. Dentro del ecosistema de Hadoop, Apache Spark [**?**] aparece como una alternativa no para de ganar popularidad. Spark es una plataforma basada en procesamiento distribuido de datos

en memoria. Existen bibliotecas de algoritmos de minería de datos implementadas sobre Hadoop y Spark, llamadas Mahout [**?**] y MLlib [**?**], respectivamente. Estas bibliotecas surgen con el objetivo de dar soporte a la tarea de la minería de datos cuando ésta es aplicada sobre conjuntos de datos masivos.

2. Hipótesis de Partida

En el momento de iniciar la tesis apenas existían en la literatura científica propuestas que aplicaran algoritmos de reducción de datos sobre conjuntos grandes, más allá de algunos basados en técnicas estadísticas simples. Las pocas propuestas encontradas no eran adecuadas debido a que: o trabajaban con bases de datos pequeñas, del orden de cientos o miles de instancias; o no eran lo suficientemente escalables para abordar problemas del orden de millones de datos.

Desde el punto de vista de la la clasificación, la aplicación de técnicas de reducción de datos en conjuntos grandes nos permitará obtener modelos más precisos y simples, así como reducir el tiempo de entrenamiento, lo cuál es especialmente relevante en Big Data. De hecho, problemas que antes no podían ser abordados (o era impracticables) debido a su gran tamaño, podrán ser abordados de manera directa gracias a las técnicas de reducción de datos distribuidas.

Surge por tanto la necesidad de desarrollar técnicas de reducción de datos distribuidas y escalables, capaces de procesar de manera eficiente grandes bases de datos y extraer conjuntos de datos de menor tamaño y mayor calidad. Como punto de partida, se establecen las siguientes hipótesis:

- Las pocas técnicas encontradas en la literatura o están basadas en técnicas estádisticas simples o no son escables.
- Con un diseño distribuido adecuado de las técnicas de reducción sería posible mantener el mismo tiempo de procesamiento frente a un aumento del tamaño de la base de datos, aumentando el número de equipos en la arquitectura subyacente.
- En general, las técnicas de reducción de datos son especialmente relevantes para grandes conjuntos de datos debido a su magnitud en las dos dimensiones de un problema: número de atributos y ejemplos. Estas técnicas nos permitirán reducir el tiempo de entrenamiento, simplificar el modelo y mejorar su precisión.
- Muchos problemas que no eran abordables o eran impracticables (millones de atributos y/o ejemplos) en el pasado podrán ser procesados eficientemente gracias a las técnicas de reducción de datos distribuidas.

Cada grupo de técnicas de reducción de datos tiene su objetivo definido y campo de aplicación específico. Sin embargo, todas son complementarias y pueden ser aplicadas a la vez en un mismo problema. Desarrollar técnicas de reducción escalables y eficientes para cada uno de estos subgrupos facilitaría la aplicación de técnicas de minería de datos sobre conjuntos de datos masivos.

3. Objetivos

En base a estas hipótesis se planteó como objetivo general el desarrollo de algoritmos de reducción de datos que den soporte al problema del procesamiento de datos de gran magnitud (Big Data), aprovechando de forma óptima y flexible los recursos disponibles. Más concretamente, en esta tesis se consideran los siguientes objetivos concretos:

- Desarrollo de nuevas propuestas algorítmicas para reducción de datos (centrándonos en selección/generación de instancias, selección/extracción de características y discretización) de manera que éstas sean fácilmente escalables a los conjuntos de datos de gran tamaño.
- Adaptación de las propuestas para Big Data por el paradigma MapReduce y para su implementación en las plataformas Hadoop y Spark (ésta última cuándo necesitemos la implementación de procesos iterativos y/o un uso intensivo de memoria).

4. Metodología y Plan de Trabajo

Dada la necesidad de una metodología teórico-práctica, se requiere un método de trabajo basado en el método científico habitual y dé cabida a las necesidades de dicha metodología. En particular, el método seguido es el siguiente:

1. **Observación**: Estudio pormenorizado del problema de la minería de datos sobre Big Data mediante el uso de técnicas de reducción escalables como etapa previa fundamental a la etapa de minería de datos. Así como, realizar un estudio de las posibilidades que ofrecen la computación distribuida y el Cloud Computing para dar solución a este problema.
2. **Formulación de hipótesis**: Diseño de nuevos algoritmos de pre-procesamiento, en concreto, de reducción de datos (discretización, selección/generación de instancias y selección/extracción de características) que usen algoritmos de nueva concepción para conseguir una reducción notable en el volumen de datos, utilizando el paradigma MapReduce.
3. **Recogida de observaciones**: Obtención de resultados como consecuencia de la aplicación de los algoritmos a bases de datos reales con volúmenes de datos importantes.
4. **Contraste de hipótesis**: Comparación de los resultados obtenidos por los algoritmos de aprendizaje sobre las bases de datos reducidas, para analizar la calidad de las propuestas. Para ello, se utilizarán de los algoritmos de las bibliotecas Mahout y Mllib.
5. **Demostración o refutación de hipótesis**: Aceptación o rechazo y modificación, si procede, de las técnicas desarrolladas como consecuencia de las pruebas realizadas.
6. **Tesis o teoría científica**: Extracción, redacción y aceptación de las conclusiones obtenidas durante el proceso.

Para alcanzar los objetivos siguiendo esta metodología, se definieron los siguientes pasos del plan de trabajo:

1. Análisis de la escalabilidad: Estudio pormenorizado del problema de la escalabilidad en la etapa de reducción de datos. Estudiando propuestas escalables en la temática y formulando nuevas ideas para la adaptación de esta etapa de la minería de datos a los nuevos paradigmas de computación distribuidos.
2. Diseño de nuevas propuestas para la obtención de algoritmos de reducción de datos (discretización, selección/generación de instancias y selección/extracción de características) más eficientes y escalables, adaptadas al procesamiento de grandes bases de datos bajo el paradigma MapReduce.
3. Implementación de las propuestas sobre plataformas de procesamiento de datos distribuido, como Hadoop y Spark.
4. Validación de los resultados obtenidos comparándolos con los de otras técnicas consideradas referentes en la investigación actual en el área mediante las herramientas estadísticas adecuadas y el uso de las bases de datos existentes actualmente.

5. Relevancia

No cabe duda de que el problema de la reducción de datos sobre conjuntos de datos masivos es de gran actualidad, y de vital importancia en diversos ámbitos (ciencia, finanzas, marketing, etc.). La propuesta de tesis pretende obtener soluciones aplicables a bases de datos del orden de cientos de miles de instancias en adelante.

A nivel científico, el trabajo desarrollado hasta el momento para esta tesis ha dado lugar a varias publicaciones en revistas internacionales. Una de ellas surge del profundo estudio realizado sobre el problema:

1. S. García, S. Ramírez-Gallego, J. Luengo, F. Herrera. Big data preprocessing: Methods and Prospects. Big Data Analytics (Submitted)

Los avances en la vertiente práctica de la tesis han dado lugar al planteamiento de seis artículos (dos publicados, tres sometidos y uno aceptado), demostrando la consecución de parte de los objetivos planteados:

1. S. Ramírez-Gallego, S. García, J. M. Benítez and F. Herrera. Multivariate Discretization Based on Evolutionary Cut Points Selection for Classification. IEEE Transactions on Cybernetics, vol. 46, no. 3, pp. 595–608, 2016.
   - Se propone un algoritmo evolutivo y multivariado para la evaluación de puntos de corte en problemas de discretización.
2. S. Ramírez-Gallego, S. García, H. Mouriño-Talin, D. Martínez-Rego, V. Bolón-Canedo, A. Alonso-Betanzos, J.M. Benitez, F. Herrera. Data Discretization: Taxonomy and Big Data Challenge. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 6, no. 1, pp. 5–21, 2016.

- Se presenta un algoritmo distribuido de discretización basado en la evaluación y selección de puntos de corte usando el principio de mínima longitud de descripción.

3. S. Ramírez-Gallego, H. Mouriño-Talín, D. Martínez-Rego, V. Bolón-Canedo, J.M. Benitez, A. Alonso-Betanzos, F. Herrera. An Information Theoretic Feature Selection Framework for Big Data under Apache Spark. IEEE Transactions on Systems, Man and Cybernetics: Systems (Submitted)
   - Se propone un conjunto de métodos distribuidos basados en teoría de la información para selección de características.
4. S. Ramírez-Gallego, I. Lastra, D. Martínez-Rego, V. Bolón-Canedo, J.M. Benitez, F. Herrera, A. Alonso-Betanzos. Fast-mRMR: Fast minimum Redundancy Maximum Relevance algorithm for high dimensional big data. Special issue on Recent Trends in Intelligent Systems ('International Journal of Intelligent Systems') (Accepted)
   - Se presenta una versión escalable del algoritmo de selección de características con varias optimizaciones importantes.
5. S. Ramírez-Gallego, S. Garcia, J.M. Benitez, F. Herrera. A Distributed Evolutionary Multivariate Discretizer for Big Data processing on Apache Spark. Special issue on Theoretical and Algorithmic Foundation for Big Data ('Journal of Computer and System Sciences') (Submitted)
   - Se propone un algoritmo de discretización distribuido y evolutivo que evalua puntos de corte usando cromosomas con formato binario y una función de fitness de tipo wrapper.
6. S. Ramírez-Gallego, B. Krawczyk, S. Garcia, M. Woźniak, J.M. Benitez, F. Herrera. Nearest Neighbor Classification for High-Speed Big Data Streams Using Spark. IEEE Transactions on Neural Networks and Learning Systems. (Submitted)
   - Se presenta un algoritmo distribuido de selección de instancias (edición) basado en vecinos cercanos para la clasificación de flujos de datos masivos.

El trabajo futuro se centra en la mejora del sistema de selección de instancias distribuido mediante la incorporación de técnicas de condensación de datos.