Parallel Data Mining Algorithms on GPUs

Alberto Cano

Department of Computer Sciences and Numerical Analysis
acano@uco.es

Abstract. La minería de datos consiste en la aplicación de algoritmos específicos para extraer conocimiento útil, comprensible y no trivial de conjuntos de datos. En los últimos años, la minería de datos se ha transformado para dar respuesta a nuevos problemas y retos cada vez más complejos sobre conjuntos de datos de creciente dimensionalidad, fenómeno conocido como *bigdata*.

Keywords: Data Mining, Big Data, Parallel Computing, GPUs

1 Introducción

Las técnicas clásicas de minería de datos ya no son capaces de obtener resultados satisfactorios en los nuevos tipos de problemas que se afrontan en la actualidad. El tamaño y la complejidad de los nuevos conjuntos de datos, los nuevos tipos de problemas que afrontar, junto con la necesidad de dar resultados exactos y en tiempo razonable, exige una nueva generación de algoritmos que sean capaces de dar respuesta a estas demandas tanto en el ámbito de la investigación científica como en el de aplicaciones empresariales y soluciones para la sociedad del siglo XXI.

La revolución tecnológica ofrece en la actualidad multitud de sistemas con alta capacidad de cómputo a bajo coste y al alcance de empresas y usuarios. Los procesadores multi-núcleo y las tarjetas de procesamiento gráfico (GPUs) inundan el mercado y ya están presentes en la mayoría de equipos en hogares, empresas y centros de investigación. Disponemos de toda una infraestructura de computación ubicua e intercomunicada lista para resolver nuestras crecientes necesidades de cómputo. En este proyecto de tesis doctoral investigaremos la aplicación de la computación en GPUs al diseño y aceleración de algoritmos de minería de datos, especialmente sobre grandes conjuntos de datos.

2 Hipótesis de Partida

La paralelización de los algoritmos en procesadores multi-núcleo ha sido una solución clásica al problema del elevado tiempo de cómputo. Sin embargo, estas soluciones se limitan habitualmente a una mera paralelización en pocos hilos, logrando una aceleración limitada. Por lo tanto, siguen sin dar solución a los verdaderos problemas y retos de computación de los algoritmos.

La computación de propósito general usando GPUs ha demostrado desde sus inicios un verdadero cambio de mentalidad en la forma de abordar la computación paralela. Ya no se trata de una mera paralelización del código, sino de un cambio radical en la forma de diseñar los algoritmos para aprovechar las enormes capacidades de cómputo de las GPUs. Problemas muy complejos o de gran tamaño son resolubles en tiempo razonable empleando millones de hilos que cooperan en la solución del problema. La gran mayoría de tipos de problemas en minería de datos son resolubles mediante algoritmos masivamente paralelos. Por lo tanto, la sinergia en el diseño de nuevos algoritmos de minería de datos masivamente paralelos bajo un paradigma de computación en GPU es un nuevo campo de investigación y aplicación con alto interés científico y comercial. Prueba de esta tendencia es el elevado número de publicaciones científicas en este campo (scopus 2012: 2726, 2011: 2441, 2010: 1795, 2009: 1217, 2008: 807). Los primeros artículos científicos publicados sobre la computación de propósito general en GPU y sus aplicaciones recibieron una gran acogida y actualmente poseen cientos de citas. Rápidamente su aplicación se ha extendido a grandes campos de la inteligencia artificial, visión artificial, metaheurísticas y minería de datos, donde numerosos autores de prestigio como W. B. Langdon han tomado la computación en GPU como un nuevo paradigma para la resolución de problemas.

3 Objetivos

El objetivo de esta tesis doctoral es el estudio, análisis, diseño, implementación y evaluación de algoritmos de minería de datos en GPUs. El desarrollo de la tesis es multidisciplinar puesto que se deberá conocer y aplicar los conocimientos relativos al diseño hardware de las GPUs, junto a la resolución e innovación en las demandas de soluciones algorítmicas a problemas de minería de datos. Específicamente, se desglosan dos grandes objetivos:

- La paralelización de algoritmos ya existentes con un elevado coste computacional para lograr su ejecución en un tiempo razonable.
- El diseño y propuesta de nuevos algoritmos, ideados ya bajo el paradigma de computación masivamente paralela en GPU, para lograr mejores resultados que algoritmos del estado del arte en las tareas de minería de datos.

4 Metodología y Plan de Trabajo

En plan de la trabajo de la tesis doctoral se enmarca bajo las líneas de actuación de la beca predoctoral perteneciente al programa de formación del profesorado universitario (FPU) del Ministerio de Educación en el periodo abarcado entre Diciembre de 2011 y Diciembre de 2015. El plan de trabajo previsto se divide de la siguiente forma:

– Primer año: revisión bibliográfica completa de la investigación desarrollada en computación en GPUs y minería de datos, incluyendo modelos de clasificación y asociación. Identificación de los retos y problemas abiertos en las técnicas de minería de datos.
– Segundo y tercer año: desarrollo e implementación de algoritmos de minería de datos en GPU. Estudios experimentales. Publicaciones científicas.
– Cuarto año: aplicación de su funcionamiento en distintos marcos de trabajo y aplicaciones reales para su transferencia al mundo empresarial y a la sociedad. Redacción definitiva de la memoria de tesis.

En cuanto a la metodología, se seguirá el método científico para lograr los objetivos mencionados:

– Establecimiento de la hipótesis: desarrollo de nuevos algoritmos en GPU para la extracción de conocimiento, mejora y adaptación de los algoritmos disponibles para problemas específicos, y análisis de los mecanismos para comparar dichos algoritmos.
– Recopilación de datos: obtención de bases de datos públicas y actuales sobre las que extraer conocimiento.
– Validar las hipótesis y las observaciones: evaluar la calidad y rendimiento de los algoritmos de minería de datos para la extracción de conocimiento respecto a las bases de datos utilizadas.
– Readaptación de la hipótesis inicial tomando en cuenta los resultados obtenidos. Esto implicará la modificación y ajuste de los algoritmos y los mecanismos de análisis acerca de su comportamiento como resultado de validaciones llevadas a cabo y de la experiencia acumulada.

5 Relevancia

El interés y la importancia del trabajo desarrollado en esta tesis doctoral será muy relevante en el ámbito científico de la minería de datos. Ofrece soluciones reales, prácticas y de bajo coste a los nuevos problemas y retos en minería de datos [1]. Los ingente cantidad de publicaciones anuales en este campo demuestra la utilidad de la computación en GPU en técnicas de inteligencia artificial que requieren altos costes computacionales, bien por la complejidad algorítmica o por el gran tamaño de los conjuntos de datos. Temas punteros hoy en día como son *big data* y *cloud computing* incluyen soluciones tecnológicas basadas en procesamiento paralelo y distribuido con GPUs.

Los resultados obtenidos en los dos primeros años de investigación han demostrado alcanzar un gran rendimiento en la aceleración de algoritmos de minería de datos ya existentes. Concretamente, hemos logrado acelerar la ejecución de algoritmos evolutivos de reglas de clasificación [2–4] y de minería de reglas de asociación [5]. Además, el uso de GPUs nos ha permitido el desarrollo de nuevos modelos masivamente paralelos [6–8], con los que hemos conseguido resultados mejores que los algoritmos del estado del arte. Recientemente, también hemos

logrado buenos resultados en la aceleración de algoritmos de reglas de clasificación de datos multi-instancia, así como en la paralelización de algoritmos de discretización de datos [9].

En la actualidad continuamos investigando y desarrollando con gran interés nuevos modelos masivamente paralelos en GPU sobre otras tareas de minería de datos [10]. Es el caso de la clasificación multi-etiqueta, donde además del problema de la dimensionalidad, existen otros problemas abiertos como el desbalanceo de las etiquetas. Además, nos estamos introduciendo en otras áreas de la inteligencia artificial, como es la visión artificial, para acelerar algoritmos de *markerless human motion capture*, donde también estamos obteniendo buenos resultados.

La presencia de GPUs en clusters y en millones de ordenadores por todo el mundo permite el desarrollo de aplicaciones de investigación y comerciales que aceleren el cómputo en GPU. De una forma totalmente transparente para el usuario, es posible convertir su ordenador en una plataforma de cómputo muy potente para la resolución de problemas de inteligencia artificial, ingeniería, biomedicina, física, química, etc. De esta forma, se logra la transferencia del desarrollo logrado por la investigación a la sociedad.

References

1. A. Cano, J. M. Luna, A. Zafra, and S. Ventura, "A Classification Module for Genetic Programming Algorithms in JCLEC," *Journal of Machine Learning Research*, vol. 16, pp. 491–494, 2015.
2. A. Cano, A. Zafra, and S. Ventura, "Speeding up the evaluation phase of GP classification algorithms on GPUs," *Soft Computing*, vol. 16, no. 2, pp. 187–202, 2012.
3. A. Cano, J. Olmo, and S. Ventura, "Parallel Multi-Objective Ant Programming for Classification Using GPUs," *Journal of Parallel and Distributed Computing*, vol. 73, no. 6, pp. 713–728, 2013.
4. A. Cano, A. Zafra, and S. Ventura, "Speeding up multiple instance learning classification rules on GPUs," *Knowledge and Information Systems*, vol. 44, no. 1, pp. 127–145, 2015.
5. A. Cano, J. M. Luna, and S. Ventura, "High Performance Evaluation of Evolutionary-Mined Association Rules on GPUs," *Journal of Supercomputing*, vol. 66, no. 3, pp. 1438–1461, 2013.
6. A. Cano, A. Zafra, and S. Ventura, "Parallel evaluation of Pittsburgh rule-based classifiers on GPUs," *Neurocomputing*, vol. 126, pp. 45–57, 2014.
7. A. Cano and S. Ventura, "GPU-parallel Subtree Interpreter for Genetic Programming," in *Proceedings of the Conference on Genetic and Evolutionary Computation*, 2014, pp. 887–894.
8. A. Cano, S. Ventura, and K. Cios, "Scalable CAIM discretization on multiple GPUs using concurrent kernels," *J. of Supercomputing*, vol. 69, no. 1, pp. 273–292, 2014.
9. A. Cano, A. Zafra, and S. Ventura, "A parallel genetic programming algorithm for classification," in *Proceedings of the 6th International Conference on Hybrid Artificial Intelligent Systems (HAIS)*, vol. 6678, no. 1, 2011, pp. 172–181.
10. A. Cano, S. Ventura, and K. Cios, "Multi-Objective Genetic Programming for Feature Extraction and Data Visualization," *Soft Computing*, vol. In press, 2015.