Doctoral Consortium: Bayesian Networks for High Dimensional and Big Data domains over new distributed computing paradigms

Jacinto Arias

Departamento de Sistemas Informáticos, Universidad de Castilla-La Mancha
jacinto.arias@uclm.es

- **Directores:** José Antonio Gámez Martín, José Miguel Puerta Callejón.
- **Año de inicio de la tesis:** 2014

1. Resumen

Durante los últimos años hemos podido observar como el impacto del fenómeno Big Data ha influenciado prácticamente todas las áreas del contexto tecnológico actual tanto por los nuevos desafíos que presenta así como por los resultados obtenidos de sus aplicaciones. En lo que respecta al campo de la Minería de Datos se imponen una serie de nuevos problemas que deben ser resueltos, no solo con respecto a la escalabilidad de las técnicas de análisis sino también por las implicaciones de la naturaleza de los datos en su manipulación.

En el aspecto tecnológico se han adoptado rápidamente una serie de herramientas que conforman los inicios de lo que podría ser un cambio de paradigma. El uso de equipos de alto rendimiento tradicionales basados en paradigmas de computación secuencial ha dejado paso a nuevas propuestas basadas en técnicas de computación paralela y con una definición cercana a la programación funcional. La principal aportación de estas herramientas con respecto al paralelismo clásico radica en su enfoque, completamente orientado al procesamiento masivo de datos, junto a su gran capacidad de abstracción que permite al programador centrarse en la algorítmica por encima de los detalles de las arquitecturas subyacentes.

El formalismo más relevante es sin duda la metodología de programación MapReduce (MR) [3], propuesta en 2004 y que ha experimentado una implantación total en el contexto tecnológico actual, principalmente gracias a su distribución y extensión por la comunidad de software libre que conforma el ecosistema de librerías software construido en torno a la plataforma abierta Apache Hadoop [12]. No obstante, y como suele ocurrir al inicio de estos periodos de innovación tecnológica, en unos pocos años han aparecido nuevas tecnologías que rivalizan con MR al solventar algunos de sus principales inconvenientes, como es el caso de Apache Spark [13], propuesto en 2012 y que ya se encuentra establecido como el paradigma de programación a utilizar en el contexto Big Data.

Como parte fundamental de estas comunidades podemos encontrar plataformas específicas, diseñadas a más alto nivel sobre Hadoop o Spark, para el

análisis de grandes volúmenes de datos mediante técnicas de Minería de Datos y Aprendizaje Automático [9,6]. Sin embargo, es común que los métodos que se proporcionan en estas distribuciones se encuentren limitados a aquellos más populares, normalmente los que mejor combinan sencillez en su definición con eficiencia y potencia analítica. La idoneidad de otros formalismos más sofisticados debe ser evaluada sobre este nuevo paradigma de manera extensa si se pretende su implantación dentro de la comunidad.

El proyecto de tesis que presentamos parte de esta situación con el objeto de estudiar las aplicaciones escalables de un formalismo concreto dentro del ámbito del Aprendizaje Automático, concretamente nos centraremos el campo de los modelos gráficos probabilísticos y en especial en el caso de las redes Bayesianas (RBs) [10]. Este formalismo ha acumulado una gran popularidad a lo largo de los años debido a sus propiedades como modelo tanto descriptivo como predictivo, además de estar basado sobre uno de los principios matemáticos más robustos como es la Teoría de la Probabilidad.

Debido a su gran complejidad la mayor parte de problemas que implican el uso de RBs presentan una dificultad NP-dura, como es el caso de el aprendizaje de los modelos o de las distintas tareas de inferencia sobre los mismos. Es por ello que la escalabilidad de dichas técnicas siempre ha sido un campo de referencia entre los investigadores de la materia [1], no obstante, las aportaciones en el dominio de Big Data son reducidas en este punto [2,7,5]. En este proyecto estudiaremos si la definición de estos modelos puede ser adaptada a sistemas con una naturaleza puramente distribuida, identificando para ello nuevos patrones y técnicas de implementación así como propiedades de este formalismo con el objeto de proponer nuevas alternativas que permitan escalar las técnicas tradicionales a dominios de gran complejidad. Los problemas a estudiar no solo presentarán retos por su elevado número de ejemplos sino también por tratarse de dominios de alta dimensionalidad, lo que en el ámbito de las RBs supone en muchos casos la inviabilidad de su aplicación.

2. Metodología y Plan de Trabajo

A la hora de desglosar el proyecto descrito en objetivos se pueden diferenciar tres etapas si bien éstas se solaparán durante el desarrollo de la tesis:

- **Formación y entorno de experimentación:** Con especial énfasis al comienzo del proyecto será necesario un periodo formativo intenso, no solo en el aspecto científico usual mediante la recopilación y el estudio del estado del arte y referencias básicas, sino en un aspecto puramente técnico. Para poder obtener los primeros resultados científicos será necesario conocer detalladamente el funcionamiento de los distintos entornos de programación distribuida así como de los sistemas de computación en los que éstos se ejecutan. Para ello será necesario partir del uso de librerías como Apache Hadoop y Apache Spark y sus subproyectos asociados y poder configurar y utilizar entornos de ejecución propios o externos.

- **Evaluación y paralelización de algoritmos existentes:** El primer paso en el proyecto será estudiar las ventajas y limitaciones de técnicas ya existentes para evaluar posibles adaptaciones a entornos distribuidos. Una vez identificados los problemas será necesario proponer y evaluar las soluciones que hayan permitido la adaptación de los algoritmos, enfatizando los principales inconvenientes que debieran ser solventados mediante el diseño de nuevas técnicas más específicas. Para ello será además necesario estudiar dominios de aplicación cuya complejidad pueda reflejar las ventajas de escalabilidad que los algoritmos distribuidos presentan frente a su alternativa secuencial. En este objetivo concreto podemos diferenciar distintas áreas que podrán ser estudiadas:
  - **Aprendizaje Automático de RBs:** Generalmente este problema se ha estudiado en dos fases diferenciadas. Por una parte, y quizás donde más se ha enfatizado los estudios de escalabilidad se encuentran las técnicas de aprendizaje estructural de los modelos, donde se trata de fijar las relaciones de dependencia entre las variables del problema mediante el aprendizaje del grafo que define el modelo. Para esta tarea las técnicas más utilizadas proponen estrategias de búsqueda heurística a partir de métricas de evaluación, el principal problema que presentan para su paralelización es una naturaleza completamente iterativa. Una vez definida la estructura del modelo es necesario estimar una serie de parámetros numéricos a partir de los datos, generalmente obteniendo distribuciones de probabilidad mediante estimación por verosimilitud a partir de los datos, en este caso, un proceso mucha más idóneo para su paralelización. Durante el proyecto se estudiarán las técnicas más prometedoras, con el fin determinar la factibilidad de las mismas a entornos distribuidos.
  - **Inferencia en RBs:** Otro problema de gran importancia en este ámbito son las distintas tareas de inferencia que se pueden realizar sobre los modelos aprendidos. Generalmente, se trata de un problema cuya complejidad aumenta de manera exponencial conforme al número de variables modeladas, lo que implica tiempos de ejecución inasumibles para dominios de alta dimensionalidad o cuando se requieren un gran número de consultas con respuestas en tiempo real. A lo largo de los años se han propuesto técnicas que restringen estructuralmente los modelos o algoritmos aproximados con el fin de reducir la complejidad de la inferencia; identificar una estrategia de inferencia paralela podría reducir en gran medida la inferencia en grandes modelos o acelerar el tiempo de respuesta.
  - **Clasificación supervisada:** Además de las tareas mencionadas anteriormente resulta relevante estudiar problemas donde los PGMs hayan sido aplicados con éxito. En el caso de la clasificación supervisada podemos encontrar algoritmos de clasificación basados en Redes Bayesianas [4] que han obtenido una gran popularidad. Estos algoritmos aprenden modelos específicos para solucionar este problema concreto simplificando así tanto el aprendizaje como la inferencia. No obstante, la mayoría de estas técnicas deben ser adaptadas para ser utilizadas en entornos

distribuidos si se pretende aplicarlas sobre volúmenes masivos de datos o dominios de alta dimensionalidad. Una de las motivaciones principales de esta tarea es la mayor disponibilidad de problemas de esta naturaleza en repositorios públicos y competiciones, como puede comprobarse en plataformas como *Kaggle*[1].

- **Nuevas propuestas de algoritmos distribuidos basados en RBs:** El estudio anterior permitirá identificar los principales inconvenientes que presenten las técnicas tradicionales al ser adaptadas a paradigmas distribuidos. Esto permitirá estudiar propuestas novedosas aprovechando la naturaleza de los nuevos paradigmas de programación y de los dominios de aplicación. El tamaño y complejidad de los problemas que podemos encontrar en Big Data suele presentar ciertas complicaciones que dificultan la mayoría de la tareas comentadas tales como el desbalanceo de los ejemplos o valores perdidos en mayor medida que los problemas tradicionales [11]. Por otra parte, la disponibilidad de tal volumen y diversidad de información permite aplicar técnicas de manipulación de los datos y preprocesado que han demostrado tener un impacto positivo a la hora de aprender determinados modelos, como por ejemplo aprendizaje de conjuntos de modelos en base a distintas muestras de los datos de entrada o proyecciones del conjunto de variables. En base a lo anterior en este objetivo final del proyecto se pretende proponer nuevos algoritmos pasados en RBs orientados a problemas de gran complejidad y compararlos con las técnicas tradicionales estudiadas anteriormente, no solo conforme a su escalabilidad sino también respecto a la calidad de sus resultados.

3. Relevancia e Impacto

Como hemos mencionado anteriormente, el número propuestas que al inicio del proyecto combinan Big Data con RBs son reducidas, aunque es de esperar que éstas crezcan durante su desarrollo ya que se trata de un campo emergente de gran interés tanto el ámbito científico como empresarial. Es por ello que las propuestas de este proyecto podrán ser contrastadas en foros de difusión específicos que ya empiezan a aparecer, probablemente también junto a otros proyectos doctorales en materias similares.

Respecto a la difusión y la transferencia de los resultados existen varias ventajas a la hora de trabajar en un campo tan emergente. Por una parte los resultados experimentales podrán ser difundidos gracias al gran esfuerzo comunitario que otros usuarios emplean en mantener y ampliar ciertas librerías software, muchas de ellas orientadas al Aprendizaje Automático. Los productos software que se desarrollen para los experimentos de este proyecto podrán diseñarse conforme a las especificaciones y guías de estilo de estas comunidades con el fin de integrarlos como parte de estas librerías software, obteniendo así una difusión y

[1] `https://www.kaggle.com`

transferencia de los resultados casi automática. Por otra parte, la cantidad de problemas abiertos disponibles crece de manera continua por medio de competiciones y repositorios abiertos a partir de los cuales se establece un banco de pruebas que permitirá interactuar a los diferentes investigadores al poder utilizar un sistema de comparación global para todos los trabajos.

Las predicciones para los próximos años [8] indican una gran demanda de trabajadores cualificados en el campo de Big Data, especialmente aquellos que combinen perfiles compatibles con análisis estadístico, Machine Learning y en definitiva aquellas materias relacionadas con el término acuñado como "Ciencia de Datos" o *Data Science*. El perfil formativo de este proyecto así como los productos que a partir de éste se puedan desarrollar encajan con las demandas estimadas de esta futura situación del mercado laboral, permitiendo así la transferencia directa de la experiencia obtenida durante el proyecto a distintos ámbitos de investigación tanto pública como privada.

Referencias

1. Arias, J., Gámez, J.A., Puerta, J.M.: Structural learning of bayesian networks via constrained hill climbing algorithms: Adjusting trade-off between efficiency and accuracy. International Journal of Intelligent Systems 30(3), 292–325 (2015)
2. Basak, A., Brinster, I., Ma, X., Mengshoel, O.J.: Accelerating bayesian network parameter learning using hadoop and mapreduce. In: Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining. pp. 101–108. ACM (2012)
3. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Communications of the ACM pp. 1–13 (2008)
4. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine learning 29(2-3), 131–163 (1997)
5. Goyal, A., Bonchi, F., Lakshmanan, L.V.: Learning influence probabilities in social networks. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 241–250. ACM (2010)
6. Kraska, T., Talwalkar, A., Duchi, J.C., Griffith, R., Franklin, M.J., Jordan, M.I.: Mlbase: A distributed machine-learning system. In: CIDR (2013)
7. Liu, B., Blasch, E., Chen, Y., Shen, D., Chen, G.: Scalable sentiment classification for big data analysis using naive bayes classifier. In: Big Data, 2013 IEEE International Conference on. pp. 99–104. IEEE (2013)
8. Manyika, J., Chui, M., Brown, B., Bughin, J.: Big data: The next frontier for innovation, competition, and productivity. Tech. Rep. June (2011)
9. Owen, S., Anil, R., Dunning, T., Friedman, E.: Mahout in action. Manning Shelter Island (2011)
10. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann (2014)
11. del Río, S., López, V., Benítez, J.M., Herrera, F.: On the use of mapreduce for imbalanced big data using random forest. Information Sciences 285, 112–137 (2014)
12. White, T.: Hadoop: The definitive guide. O'Reilly Media, Inc. (2012)
13. Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I.: Spark: cluster computing with working sets. In: Proceedings of the 2nd USENIX conference on Hot topics in cloud computing. pp. 10–10 (2010)