Big Data: Diseño de algoritmos para clasificación extremadamente no balanceada

Sara Del Río García
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.
srio@decsai.ugr.es
Fecha de Inicio: Octubre de 2013

Palabras clave: Big Data, Clasificación, Preprocesamiento, Datos Desbalanceados, Sistemas Distribuidos, Minería de Datos

1. Introducción

En la actualidad, el análisis e interpretación de grandes volúmenes de datos representa una necesidad fundamental pues la extracción de conocimiento a partir de los mismos puede ayudar a las organizaciones a enfrentarse a nuevos problemas o desafíos. Como solución al problema de análisis en grandes bases de datos aparecieron los algoritmos de extracción de conocimiento (*Knowledge Discovery in Databases*) o minería de datos (en inglés, *Data Mining*) [1].

Con los avances de tecnología en diversas áreas tales como sistemas de sensores, comunicaciones o almacenamiento, la cantidad de datos que se están generando es cada vez mayor, tanto que la gran parte de los datos en el mundo se han generado recientemente. Estas enormes cantidades de información incluyen diferentes tipos de datos (estructurados/no estructurados), diferentes tamaños (desde terabytes hasta zettabytes) y pueden provenir de multitud de sectores como el de las telecomunicaciones, el farmacéutico o el de la salud. Estos datos se conocen como grandes bases de datos (en inglés, *big data*) [2].

La extracción de conocimiento a partir de *big data* es considerado un problema importante para la obtención de información útil ya que los ordenadores actuales no pueden manejar dicha información de manera sencilla. Por este motivo, es cada vez más importante el desarrollo de herramientas que permitan el análisis y la interpretación de tales cantidades de datos para la extracción de conocimiento interesante a partir de las operaciones actuales de las organizaciones y con el fin de prever ciertas operaciones críticas. Además, éste creciente aumento de los datos supone un desafío para los algoritmos de minería de datos y aprendizaje automático, que no pueden escalar fácilmente problemas de *big data*. De esta forma, es necesario rediseñar dichos algoritmos de modo que puedan ser aplicados a problemas del mundo real.

Una de las soluciones más populares para abordar el problema de *big data* es el modelo de programación denominado *MapReduce* [3]. Se trata de un paradigma computacional que fue presentado por Google en 2004 para el desarrollo aplicaciones distribuidas, escalables y confiables. Este nuevo paradigma consta de dos funciones principales: *Map* y *Reduce*. En términos generales, en la fase *Map* los datos se dividen en conjuntos más pequeños que son distribuidos y procesados en paralelo. A continuación, en la fase *Reduce* se combinan los resultados obtenidos en la fase anterior para producir la salida final. *Hadoop* [3] es la implementación de código abierto más popular de *MapReduce*

2. Hipótesis de Partida

Una de las cuestiones que dificulta la extracción de conocimiento es el problema de clasificación sobre conjuntos de datos desbalanceados [4]. Esta situación está presente en numerosas aplicaciones del mundo real y se produce cuando una o más clases están representadas por un gran número de ejemplos, mientras que el resto de las clases por tan sólo unos pocos. En estos problemas, el interés de los expertos se centra en la identificación de las clases menos representadas ya que suelen ser las más importantes desde el punto de vista del aprendizaje y conllevan altos costes cuando su identificación no se lleva a cabo adecuadamente.

Un factor que influye negativamente en la clasificación con conjuntos de datos desbalanceadoses es la presencia de pequeños disyuntos (en inglés, *small disjuncts*). Este problema se produce cuando los datos de una única clase están concentrados en un pequeño espacio del problema rodeados por ejemplos de la clase contraria. Este tipo de regiones son difíciles de detectar para muchos algoritmos de aprendizaje [5].

Podemos encontrar diferentes técnicas para abordar el problema del no balanceo, tales como los enfoques sensibles al coste [6], algoritmos específicos o métodos de muestreo tales como sobremuestreo o submuestreo [7]. También se encuentran las técnicas basadas en algoritmos que generan datos sintéticos o artificiales para la clase minoritaria. SMOTE (*Syntetic Minority Over-sampling Technique*) [8] es el algoritmo más conocido en éste ámbito.

Cuando nos centramos en el ámbito de *big data*, nos planteamos la extensión de los algoritmos actuales bajo el paradigma *MapReduce*. La extensión de los enfoques sensibles al coste y de muestreo son fáciles de plantear. Sin embargo, la extensión directa de SMOTE no es una solución por cuanto el comportamiento es bastante malo en comparación con los algoritmos de muestreo. Por este motivo, el diseño de nuevos algoritmos para la generación de datos artificiales en el contexto de problemas de *big data* desbalanceados supone un gran desafío.

Desde la perspectiva de los algoritmos de aprendizaje, y en particular de los algoritmos de aprendizaje de reglas, hemos de destacar que el aprendizaje basado en reglas es una de las principales aproximaciones en aprendizaje automático [9]. Éstas tecnologías proporcionan un amplio conjunto de algoritmos de aprendizaje. Su principal objetivo consiste en descubrir relaciones interesantes que puedan ayudar a entender mejor las dependencias entre diferentes variables en bases de datos y que estas relaciones representadas en reglas nos permitan diseñar un sistema de clasificación. A lo largo de los años se han desarrollado numerosos

Sistemas Basados en Reglas con el fin de hacer frente a los problemas de clasificación. Estos sistemas han sido utilizados con éxito en numerosas aplicaciones debido a la compacidad de la representación del conocimiento descubierto, a la accionabilidad de las reglas aprendidas y a su interpretabilidad. Sin embargo, con el fin de hacer frente a *big data*, los enfoques clásicos de aprendizaje de reglas deben ser rediseñados y adaptados mediante el diseño de procedimientos de fusión de reglas en la fase *Reduce* dentro de un esquema *MapReduce*.

El aprendizaje basado en combinación de clasificadores es una de las áreas más prometedoras en aprendizaje automático que ha demostrado un buen comportamiento en muchas aplicaciones del mundo real. Estos enfoques construyen un conjunto de clasificadores para después clasificar los datos mediante la votación de sus predicciones. Dos de los enfoques más representativos del aprendizaje basado en combinación de clasificadores son *bagging* [10] y *boosting*. Una cuestión importante en estos enfoques es la técnica para combinar las predicciones (o esquema de votación) de los clasificadores para *big data*, ya que pueden dar resultados distintos en función de diferentes factores. Por otro lado, la forma de dividir el conjunto de datos original puede ser importante a fin de obtener modelos más precisos basados en combinación de clasificadores. Por ello es necesario tanto desarrollar nuevos modelos de votación apropiados en la fase *Reduce*, como diseñar nuevos mecanismos de división de datos en la fase *Map* para algoritmos basados en combinación de clasificadores en el escenario de *big data*.

Finalmente, hemos de destacar un importante problema que surge al aplicar el pardigma *MapReduce* al problema de clasificación no balanceada. La forma secuencial de dividir el conjunto de datos en bloques puede provocar la aparicion de *small disjuncts* [5]. Esta puede ser una de las causas del mal comportamiento de las técnicas de preprocesamiento que crean instancias artificiales, como SMOTE. Por ello, será necesario diseñar algoritmos de preprocesamiento que permitan abordar dicho problema.

3. Objetivos

El objetivo general en esta tesis se centra en el desarrollo de algoritmos desde una doble perspectiva: (1) para hacer frente a los problemas de clasificación con *big data* en el contexto de los Sistemas Basados en Reglas y de los algoritmos basados en combinación de clasificadores desde el punto de vista del modelo; (2) y para abordar problemas de *big data* desbalanceados desde el punto de vista de los datos; considerando el aprendizaje de reglas para problemas de *big data* desbalanceados. Más concretamente, se consideran los siguientes objetivos:

1. Desarrollo de algoritmos para abordar problemas de *big data* desbalanceados usando el paradigma de programación *MapReduce*.
2. Diseño de estrategias de combinación de reglas en la fase *Reduce* de un proceso *MapReduce* para Sistemas Basados en Reglas y, diseño de nuevos esquemas de votación en la fase *Reduce* para algoritmos basados en combinación de clasificadores en el escenario de *big data*.
3. Diseño de nuevos mecanismos para división de datos en la fase *Map* y, diseño de algoritmos que permitan abordar los *small disjuncts* en problemas de *big data* desbalanceados usando el paradigma de programación *MapReduce*.

4. Metodología y Plan de Trabajo

Para el desarrollo de los objetivos se ha seguido el método científico tradicional, cuyas etapas se describen a continuación:

1. **Observación**: Estudio pormenorizado del problema de minería de datos sobre *big data*.
2. **Formulación de hipótesis**: Consiste en el desarrollo de nuevos algoritmos para clasificación con *big data* desde una doble perspectiva: (1) para hacer frente a los problemas de clasificación con *big data* en el contexto de los sistemas basados en reglas y modelos basados en combinación de clasificadores desde el punto de vista del modelo y; (2) para abordar problemas de *big data* desbalanceados desde el punto de vista de los datos.
3. **Recogida de observaciones**: Esta etapa requiere el uso de grandes bases de datos para validar las distintas propuestas presentadas.
4. **Contraste de hipótesis**: Teniendo en cuenta las observaciones de la etapa anterior, en esta etapa evaluaremos la calidad de los modelos.
5. **Demostración o refutación de la hipótesis**: aceptación o rechazo y modificación, si procede, de las técnicas desarrolladas como consecuencia de las conclusiones extraídas a partir de los estudios realizados.
6. **Tesis o teoría científica**: extracción, redacción y aceptación de las conclusiones obtenidas durante el proceso.

5. Relevancia

En la actualidad, el análisis y la interpretación de grandes volúmenes de datos representa una necesidad fundamental para la extracción de conocimiento útil y valioso para nuestro entorno socioeconómico. Con los avances de la tecnología la cantidad de datos que se está generando y almacenando es cada vez mayor. Esta gran cantidad de datos y las tecnologías que los procesan es el área que se conoce con el nombre de "big data". Esta tesis se basa en el desarrollo de tecnologías para abordar problemas de *big data* que puedan ser aplicables a nuestro entorno socioeconómico.

A nivel científico, el trabajo realizado hasta la fecha ha dado lugar a varias publicaciones en revistas internacionales. Una publicación que recoge el estado del arte es la siguiente:

1. A. Fernandez, S. Río, V. López, A. Bawakid, M.J. del Jesus, J.M. Benítez, F. Herrera, Big Data with Cloud Computing: An Insight on the Computing Environment, MapReduce and Programming Frameworks, *WIREs Data Mining and Knowledge Discovery*, 4:5 (2014) 380-409. **Contenido**: se presenta un estado del arte sobre *big data*.

El resto de publicaciones fruto de los avances en la vertiente práctica de la tesis:

1. S. Río, V. López, J.M. Benítez, F. Herrera, On the use of MapReduce for Imbalanced Big Data using Random Forest. *Information Sciences*, 285 (2014) 112-137. **Contenido**: se presentan varias técnicas tales como sobremuestreo, bajomuestreo o aprendizaje sensible al coste, adaptadas para abordar problemas de *big data* desbalanceados usando *MapReduce*.

2. S. Río, V. López, J.M. Benítez, F. Herrera, A MapReduce Approach to Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules, *International Journal of Computational Intelligence Systems*, 8:3 (2015) 422-437. **Contenido**: se presenta un sistema de clasificación basado en reglas difusas desarrollado en dos versiones con diferentes procesos de fusión de reglas para problemas de *big data*.
3. V. López, S. Río, J.M. Benítez, F. Herrera, Cost-Sensitive Linguistic Fuzzy Rule Based Classification Systems under the MapReduce Framework for Imbalanced Big Data. *Fuzzy Sets and Systems*, 258 (2015) 5-38. **Contenido**: se propone un sistema de clasificación basado en reglas difusas para problemas de *big data* desbalanceados.
4. I. Triguero, S. Río, V. López, J. Bacardit, J.M. Benítez, F. Herrera, ROSEFW-RF: The winner algorithm for the ECBDL'14 Big Data Competition: An extremely imbalanced big data bioinformatics problema, *Knowledge-Based Systems*, 87 (2015) 69-79. **Contenido**: se describe el algoritmo que ganó la competición *ECBDL'14 Big Data Competition*.
5. D. Galpert, S. Río, F. Herrera, E. Ancede, A. Antunes, G. Agüero-Chapin, An Effective Big Data Supervised Imbalanced Classification Approach for Ortholog Detection in Related Yeast Species, *BioMed Research International*, in press (2015). **Contenido**: se propone un esquema para la detección de genes ortólogos en problemas de *big data*.

Premio: se ganó la competición *ECBDL'14 Evolutionary Computation for Big Data and Big Learning*, celebrada en Vancouver (Canadá) del 12 al 16 de julio. Para ello se hizo uso del algortimo *ROSEFW-RF*.

Como trabajo futuro, se esperan nuevos avances en las líneas de fusión de clasificadores y en el preprocesamiento de datos para abordar los problemas discutidos.

Referencias

1. Han, J., Kamber, M.: (Eds.), Data mining. Concepts and techniques. Morgan Kaufmann (2011)
2. Minelli, M., Chambers, M., Dhiraj, A.: Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses. John Wiley & Sons (2013)
3. White, T.: Hadoop, The Definitive Guide. O'Reilly Media, Inc. (2012)
4. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences. 250, 113–141 (2013)
5. Jo, T., Japkowicz, N.: Class imbalances versus small disjuncts. SIGKDD Explorations 6, 40–49 (2004)
6. Elkan, C.: The foundations of cost–sensitive learning. Proceedings of the 17th IEEE International Joint Conference on Artificial Intelligence (IJCAI'01), 973–978 (2001)
7. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explorations 6, 20–29 (2004)
8. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligent Research 16, 321–357 (2002)

9. Fürnkranz, J., Gamberger, D., Lavrac, N.: Foundations of Rule Learning. Springer (2012)
10. Breiman, L.: Bagging predictors. Machine Learning 24, 123–140 (1996)