Científicos de la UGR ganan un importante concurso internacional de informática sobre ‘big data’

04/09/2014

Pertenecen al grupo de investigación “Soft Computing y Sistemas de Información Inteligentes”, que dirige el catedrático de la Universidad de Granada Francisco Herrera

Los “big data” son conjuntos de datos de un elevado tamaño cuyo volumen, diversidad y complejidad requieren el uso de nuevas arquitecturas, técnicas, algoritmos y análisis para gestionar y extraer el valor y conocimiento oculto en ellos

Científicos de la Universidad de Granada, pertenecientes al grupo de investigación “Soft Computing y Sistemas de Información Inteligentes” (SCI2S), han ganado la “ECBDL’14 Big Data Competition”, un concurso celebrado este verano en Vancouver (Canadá), en el marco del congreso internacional GECCO-2014.

Este certamen, uno de los más prestigiosos del mundo en este ámbito de investigación, premia los mejores trabajos relacionados con los “big data”, conjuntos de datos de un elevado tamaño cuyo volumen, diversidad y complejidad requieren el uso de nuevas arquitecturas, técnicas, algoritmos y análisis para gestionar y extraer el valor y conocimiento oculto en ellos.

La “ECBDL’14 Big Data Competition” se ha centrado en esta edición en un problema de clasificación en bioinformática. En concreto, los participantes debían trabajar sobre un conjunto de datos del campo de la predicción de estructuras de proteínas, en el que se pretendía conseguir un predictor para distinguir un conjunto de estructuras a partir de las ya conocidas, especialmente la detección de contactos residuo-residuo en las proteínas.

El conjunto de entrenamiento utilizado en la competición constaba de dos clases, con alrededor de 32 millones de instancias con 631 atributos ocupando 56,7 Gigabytes de datos. Para validar la utilidad de los métodos de la competición se ha considerado un conjunto de test con unos 2,8 millones de ejemplos que se almacenan aproximadamente en 5 Gigabytes de datos.

El equipo de la UGR que ha ganado la competición ha propuesto una combinación de técnicas de preprocesamiento de datos (sobremuestreo de alta ratio sobre la clase minoritaria y selección de características basada en pesos) y multiclasificadores basados en árboles de decisión utilizando MapReduce, extendiendo las ideas publicadas en la revista “Information Sciences”. En segundo lugar quedó la Universidad de Newcastle (Reino Unido), y en tercero la Universidad de Nueva Gales del Sur (Australia).

Como explica el director del grupo de investigación “Soft Computing y Sistemas de Información Inteligentes” de la UGR, Francisco Herrera, “los desarrollos tecnológicos en torno al “big data” y el análisis inteligente de datos han dado lugar recientemente al término de Ciencia de Datos (Data Science), definido como un área emergente de trabajo relacionada con la preparación, análisis, visualización, gestión y mantenimiento de grandes colecciones de datos para la obtención de conocimiento que genere ventajas de negocio. Debido al impacto que estas temáticas están llegando a alcanzar, ha aparecido un nuevo término profesional: el “científico de datos”.

El alto potencial del “big data” ha sido reconocido de inmediato debido a su influencia sobre problemas de diversos campos de conocimiento. “Entender la economía global, obtener una mejor planificación de servicios públicos, desarrollar investigaciones científicas o buscar nuevas oportunidades de negocio son algunas de las grandes aplicaciones relacionadas con estos grandes repositorios de datos”, apunta el profesor Herrera.

Dos artículos importantes

El grupo de investigación SCI2S de la Universidad de Granada ha desarrollado diversas aproximaciones basadas en MapReduce y las tecnologías Hadoop y Spark para abordar problemas de “big data”. Estas aproximaciones tratan de lidiar con grandes conjuntos de datos, con datos heterogéneos y con datos textuales como los disponibles en las redes sociales.

Recientemente ha publicado dos trabajos en los que se aborda el problema del desbalanceo entre clases en “big data”, un problema recurrente en aplicaciones del mundo real en el que tenemos pocas instancias asociadas a un hecho concreto

frente a las muchas instancias en el problema, por ejemplo, los casos de fraude respecto al número total de transacciones.

Así, en un primer trabajo han desarrollado sistemas de clasificación basados en reglas difusas combinados con aproximaciones sensibles al coste utilizando MapReduce. Estos avances han sido publicados en la revista “Fuzzy Sets and Systems”, y se caracterizan por proporcionar clasificadores en forma de reglas con etiquetas lingüísticas, de manera que sean interpretables por el usuario y que a su vez son capaces de obtener una alta efectividad en la clasificación.

Por otra parte, en un segundo trabajo los investigadores de la UGR han estudiado la aplicación de multiclasificadores siguiendo el modelo Random Forest junto a algoritmos de preprocesamiento bajo el paradigma MapReduce, habiéndose publicado estos resultados en la revista internacional “Information Sciences”. Para abordar el desequilibrio de clases con éxito, se proponen diversas estrategias como las técnicas sensibles al coste y el uso de técnicas de preprocesamiento basadas en el muestreo de clases para tratar de obtener una distribución de instancias equilibrada que permite mejorar el funcionamiento de los algoritmos de aprendizaje.

Además, en el grupo de investigación se ha iniciado una línea de trabajo en el área conocida como “Social Big Data” para desarrollar algoritmos cuyo objetivo sea el procesamiento de información textual, como la obtenida en las redes sociales.

Imágenes adjuntas:

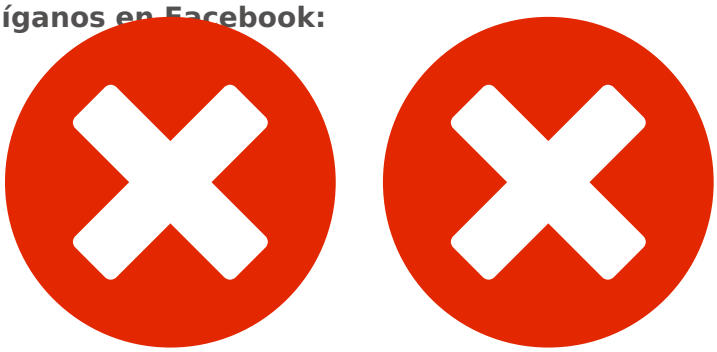
FOTO 1: José Manuel Benítez (izquierda), recogiendo el Premio en Vancouver

FOTO 2: En la foto del equipo aparecen los investigadores participantes en la competición. De izquierda a derecha: Sara del Río, Isaac Triguero, Victoria López, Francisco Herrera y José Manuel Benítez. //

Contacto:

Francisco Herrera
Director del grupo de investigación “Soft Computing y Sistemas de Información Inteligentes”
Dpto. de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada.
Tlfno: 958 240 598
Correo electrónico:
LINK: --LOGIN--d4015d6400b2cbe1b74cbbb295263f5cdecsai[dot]ugr[dot]es -> --LOGIN--d4015d6400b2cbe1b74cbbb295263f5cdecsai%5Bdot%5Dugr%5Bdot%5Des

Síganos en Facebook:



Síganos en Twitter:

- LINK: PROPUESTA DE ACTIVIDADES CANAL UGR -> http://canal.ugr.es/prensa-y-comunicacion/item/54050
- CANALUGR: RECURSOS DE COMUNICACIÓN E INFORMACIÓN
- PUBLICITE SU CONGRESO UGR
- VER MÁS NOTICIAS DE LA UGR
- BUSCAR OTRAS NOTICIAS E INFORMACIONES DE LA UGR PUBLICADAS Y/O RECOGIDAS POR EL GABINETE DE COMUNICACIÓN
- RESUMEN DE MEDIOS IMPRESOS DE LA UGR

- RESUMEN DE MEDIOS DIGITALES DE LA UGR
- RECOMENDACIONES PARA EL USO DE LAS LISTAS DE DISTRIBUCIÓN DE LA UGR
- LINK: Perfiles oficiales institucionales de la UGR en las redes sociales virtuales Tuenti, Facebook, Twitter y YouTube -> /tablon/*/boletines-canal-ugr/formulario-de-propuesta-de-actividades

Gabinete de Comunicación - Secretaría General
UNIVERSIDAD DE GRANADA
Acera de San Ildefonso, s/n. 18071. Granada (España)
Tel. 958 243063 - 958 244278
Correo e. LINK: --LOGIN--61dab3f5145154c15507d4098f0f1b4eugr[dot]es -> --LOGIN--61dab3f5145154c15507d4098f0f1b4eugr%5Bdot%5Des
Web: http://canal.ugr.es Facebook UGR Informa: https://www.facebook.com/UGRinforma
Facebook UGR Divulga: https://www.facebook.com/UGRdivulga
Twitter UGR Divulga: https://twitter.com/UGRdivulga