



Minería de Datos a Datos Masivos, con Weka sobre Hadoop, Caso de estudio: Catálogo de Galaxias

Jimenez Joaquín(1), Klenzi Raúl, Malberti Alejandra(2), Georgina Coldwell(3)

(1) Becario de Investigación alumno avanzado CICITCA-UNSJ 2013-2014

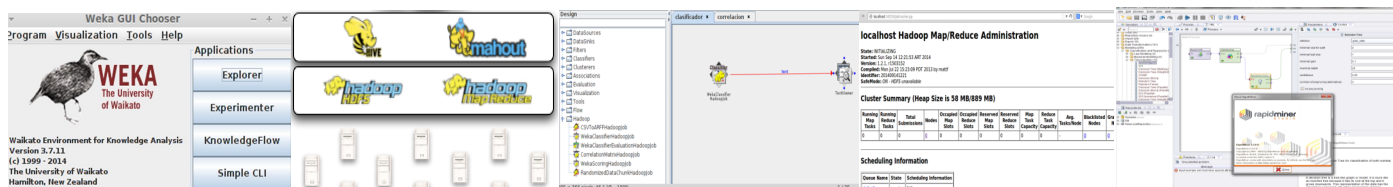
Licenciatura en Ciencias de la Computación. Dpto Informática - FCEfYN - Universidad Nacional de San Juan (UNSJ)

(2)Dpto Informática - FCEfYN - Universidad Nacional de San Juan (UNSJ) (3) Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)

Joaquinjimenez33,rauloscarklenzi,amalberti,georginacoldwell@gmail.com

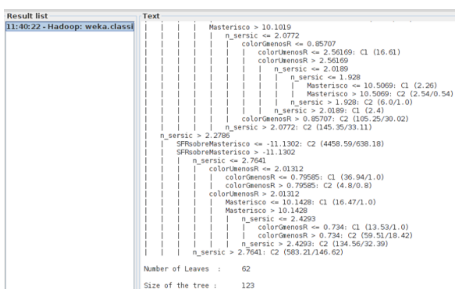
Dada la creciente necesidad de procesamiento eficiente de grandes volúmenes de datos, no alcanza con utilizar algoritmos secuenciales óptimos, sino que necesariamente se debe trabajar en paralelo. Es por esto que muchos entornos de software del área del aprendizaje de máquina y minería de datos han desarrollado extensiones para esta problemática. En particular Weka (Waikato Environment for Knowledge Analysis) ha desarrollado un plugin para ejecutar sobre Hadoop, un conocido Framework que soporta aplicaciones distribuidas bajo licencia libre, y que puede trabajar con miles de máquinas. El presente artículo muestra el uso de los plugins "distributedWekaBase" y "distributedWekaHadoop", desarrollados para la integración de Weka con Hadoop. En particular se trabaja con datos provenientes del campo de la Astronomía, un área que se destaca, entre otras cosas, por generar y dominar grandes volúmenes de información. Los algoritmos elegidos son el de obtención de la matriz de correlación y el de clasificación basada en árboles J48. Este trabajo se desarrolla en el marco de una beca de investigación para alumno avanzado dentro del proyecto bianual CICITCA-UNSJ "Extracción de Conocimiento en Datos Masivos"

Herramientas de Software Libre de Minería de Datos y Trabajo en Paralelo

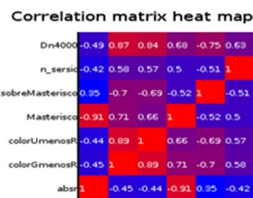


Apache Hadoop es un framework de código libre que soporta aplicaciones distribuidas sobre datos masivos. Este Framework puede ser configurado para trabajar con miles de máquinas y petabytes de información. Una importante característica de Hadoop es la división de los datos y de la computación entre muchos nodos, y de esta forma puede correr procesos paralelos en las que cada host trabaja con su porción de los datos.

Weka es una herramienta de código abierto que contiene una colección de algoritmos de aprendizaje de máquina implementados en Java. Weka desarrolló 2 plugins para integrarse con Hadoop. distributedWekaBase y distributedWekaHadoop.



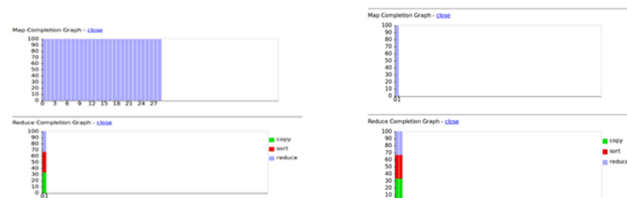
J48 es una implementación del algoritmo C4.5. Este aplica el concepto de entropía de información y construye árboles de decisión desde datos de entrenamiento. Estos datos de son un grupo de ejemplos ya clasificados, en los que cada ejemplo es un vector de atributos o características. Además hay un atributo especial C, llamado atributo de clase, que representa al tipo o familia a la que pertenece el ejemplo.



Matriz de correlación

Una matriz de correlación, es una tabla de doble entrada que muestra la misma lista multi-variable horizontal y verticalmente. Cada entrada contiene el correspondiente coeficiente de correlación, llamado r, o la relación entre cada pareja de variables, expresada por medio de un número entre 0 y 1. El modelo mide la interdependencia en relaciones asociadas entre cada pareja de variables.

Árbol de clasificación J48



Gráficas resumen de las tareas de mapeo y reducción J48

Resumen de las tareas de mapeo y reducción Matriz de correlación



Detalles de ejecución de matriz de correlación

Bibliografía recomendada

