Hibridización de K-Means a través de Técnicas Metaheurísticas

Andrea Villagra, Daniel Pandolfi
Universidad Nacional de la Patagonia Austral - Unidad Académica Caleta Olivia
Ruta 3 Acceso Norte s/n
(9011) Caleta Olivia - Santa Cruz - Argentina
{avillagra,dpandolfi}@uaco.unpa.edu.ar

and

Guillermo Leguizamón
Universidad Nacional de San Luis,
Ejército de los Andes 950, (5700) San Luis, Argentina
legui@unsl.edu.ar

Resumen

En los últimos años, ha existido un gran crecimiento en nuestras capacidades de generar y colectar datos, debido básicamente al gran poder de procesamiento de las máquinas y a su bajo costo de almacenamiento. Sin embargo, dentro de estas enormes masas de datos existe una gran cantidad de información "oculta", de gran importancia estratégica, a la que no se puede acceder por las técnicas clásicas de recuperación de la información. La Minería de Datos implica "escabar"en esa inmensidad de datos, en búsqueda de patrones, asociaciones o predicciones que permitan transformar esa maraña de datos en información útil. Una de las tareas utilizadas en minería de datos es el *clustering* (agrupamiento) y un algoritmo muy popular y simple usado en esta tarea es *K-means*. A pesar de su popularidad el mencionado algoritmo sufre de algunas dificultades. *K-means* requiere varias iteraciones sobre todo el conjunto de datos, lo cual puede hacerlo muy costoso computacionalmente cuando se lo aplica a grandes bases de datos, el número de *clusters* K debe ser suministrado por el usuario y la búsqueda es propensa a quedar atrapada en mínimos locales.
Se pretende a través de esta línea de investigación desarrollar técnicas avanzadas o mejoradas de minería de datos, particularmente en la tarea de *clustering* y además, proponer mejoras al algoritmo de *K-means* basándose en la aplicación de técnicas Metaheurísticas.

1. INTRODUCCIÓN

Factores como el avance tecnológico asociado al continuo abaratamiento de los costos, implica que los volumenes de datos almacenados crece exponencialmente. En la actualidad, estamos en una etapa en la que no es fácil visualizar los datos que están almacenados. Existen muchos dominios en los cuales la acumulación de datos es altísima y por consiguiente se hace cada vez más difícil poder obetener información relevante para la toma de decisiones basados en dichos datos.

La tarea de Minería de Datos implica "escabar"en esa inmensidad de datos, usualmente medidos en gigabytes, en búsqueda de patrones, asociaciones o predicciones que permitan transformar esa maraña de datos en información útil. La tarea de minería de datos no siempre parte de un conocimiento previo de lo que se busca en el conjunto de datos disponibles, por el contrario, es muy frecuente que no sepamos de antemano lo que se busca. Es decir, se realiza una búsqueda de patrones desconociendo el patrón que pueda surgir.

La minería de datos constituye el núcleo del análisis inteligente de los datos y ha recibido un gran impulso en los últimos tiempos motivado por distintas causas: a) el desarrollo de algoritmos eficientes y robustos para el procesamiento de grandes volúmenes de datos, b) un poder computacional más barato que permite utilizar métodos computacionalmente intensivos, y c) las ventajas comerciales y científicas que han brindado este tipo de técnicas en las más diversas áreas. Entre las áreas donde han sido utilizadas exitosamente las técnicas de minería de datos podemos mencionar distintas aplicaciones financieras y bancarias, análisis de mercado, seguros y salud privada, educación, procesos industriales, medicina, biología, bioingeniería, telecomunicaciones, Internet, turismo, deportes, etc.

2. CLUSTERING Y K-MEANS

El *Clustering* ha sido aplicado exitosamente en una amplia variedad de disciplinas científicas y de ingeniería tales como psicología, biología, medicina, vision computarizada y comunicaciones. El *clustering* organiza los datos (un conjunto de patrones, cada patrón puede ser un vector de mediciones) extrayendo estructuras subyacentes. El agrupamiento finaliza cuando los patrones dentro de un grupo (*cluster*) son más similares entre sí que con otros patrones que pertenecen a otros grupos diferentes. Por lo tanto, organizar los datos usando *clustering* emplea alguna medida de disimilitud entre los conjuntos de patrones. La medida de disimilitud se define en base a los datos bajo análisis y del propósito del análisis. Se han propuesto diferentes algoritmos de *clustering* adecuados a diversos requerimientos. Los algoritmos de *clustering* pueden clasificarse en general en jerárquicos y particionados basados en la estructura de extracción. Los algoritmos de clustering jerárquico construyen particiones de una jerarquía representadas en un dendrograma en el cual cada partición se anida con otra partición en el siguiente nivel de la jerarquía. Los algoritmos de *clustering* particionado generan una sola partición simple, con un número especificado o estimado de *clusters* no solapados, de los datos intentando recuperar grupos naturales presentes en ellos.

Uno de los problemas importantes en *clustering* particionado es encontrar una partición de los datos, con un número especificado de *clusters* que minimice la variación total dentro de los *clusters*. En general los algoritmos de *clustering* particionado son iterativos y *hill clumbing* y usalmente convergen a mínimos locales.

El algoritmo de *clustering* más simple y popular entre los algoritmos de clustering es el algoritmo de *K-means*. Dado un conjunto P, el mencionado algoritmo, conocido como algoritmo de Lloyd [13], trata de encontrar k centroides en el espacio minimizando el costo, que es la suma del cuadrado de la distancia Euclidean de cada punto en P a su centro más cercano. Al comienzo del algoritmo, se eligen de forma aleatoria del conjunto de datos originales k centroides iniciales. Luego el algoritmo se mantiene invocando *k-means* iteraciones. Cada *k-means* iteración consiste de dos operaciones. Primero, a cada punto dentro del conjunto de datos se lo asigna al centroide más cercano. Segundo, los puntos se dividen en k grupos de acuerdo al centroide más cercano en el paso previo y los centros geométricos (centroides) de todos los grupos forman un nuevo conjunto de centroides. Este procedimiento continúa hasta que los centroides se mantengan sin cambios. El algoritmo de *k-means* se usa en varias aplicaciones diferentes debido a su simplicidad y eficiencia. Sin embargo, hay tres problemas principales con el algoritmo de *k-means*. Primero, en cada iteración, se consume mucho tiempo de computación asignando a cada punto del conjunto de datos a su nuevo centroide más cercano. Segundo, el número de k centroides iniciales debe ser suministrado por el usuario. Tercero, este algoritmo puede quedar facilmente atrapado en un óptimo local. Para abordar estos problemas exiten varios trabajos que intentan acelerar la búsqueda del centroide más cercano y la elección de los centroides iniciales [8], [14], [10], [15].

3. LINEA DE INVESTIGACION

En los últimos años el grupo de investigación se enfocó en el desarrollo y conocimiento de los diferentes enfoques relacionados al campo de la inteligencia computacional, en particular al de computación evolutiva [12], [9], [11] y sus aplicaciones en la industria [4], [3], [2]. Simultáneamente, ha surgido un gran número de enfoques metaheurísticos [1], [7], [6], muchos de ellos bio-inspirados, los que a pesar de ciertas diferencias conceptuales en su diseño, comparten muchos aspectos que permiten entre otras cosas: a) aplicar conceptos que originalmente fueran diseñados para otra heurística o metaheurística con el objetivo de lograr mejoras substanciales, b) diseñar enfoques híbridos que aprovechen las ventajas relativas de cada enfoque involucrado, c) incorporar criterios de búsqueda más avanzados. Por esta razón una de las líneas de investigación dentro del Laboratorio de Tecnologías Emergentes (LabTEm) es la aplicabilidad de metaheurísticas en problemas de minería de datos y en particular en la tarea de *clustering*, donde se ha comenzado a trabajar [5]. Actualmente se esta trabajando en el desarrollo de técnicas avanzadas o mejoradas de minería de datos, particularmente en la tarea de *clustering* y en mejoras al algoritmo de *K-means* basándose en la aplicación de técnicas Metaheurísticas. Se pretende hibridizar dicho algoritmo a través de técnicas Metaheurísticas y comparar los resultados con los obtenidos en mejoras alternativas a dicho algoritmo propuestas en trabajos existentes, incluyendo además, la aplicación de los resultados a distintos problemas del mundo real analizando la calidad de dichas técnicas.

4. AGRADECIMIENTOS

EL primer y segundo autor agradecen a la Universidad Nacional de la Patagonia Austral por su apoyo al grupo de investigación y además, la cooperación de los integrantes del proyecto que continuamente proveen de nuevas ideas y críticas constructivas. El tercer autor agradece el constante apoyo brindado por la Universidad Nacional de San Luis y la ANPYCIT que financian sus actuales investigaciones.

REFERENCIAS

[1] Hussein A., Ruhul A. S., and Charles S. N. *DATA MINING: A heuristic approach*. Idea Group Publishing, 2002.

[2] Villagra A., Montenegro C., de San Pedro M., Lasso M., and Pandolfi D. Planificación con restricciones del mantenimiento de locaciones petroleras. In *XII RPIC - Reunión de Trabajo en Procesamiento de la Información y Control*, 2007.

[3] Villagra A., Montenegro C., de San Pedro M., Lasso M., and Pandolfi D. Restricciones en la replanificación del mantenimiento de locaciones petroleras. In *Congreso Argentino de Ciencias de la Computación*, 2007.

[4] Villagra A., Montenegro C., de San Pedro M., Lasso M., Vidal P., and Pandolfi D. *Mantenimiento de locaciones petroleras mediante un Algoritmo Multirecombinativo*, chapter 11, pages 319–322. CAIP, 2007.

[5] Villagra A., Pandolfi D., and Leguizamón G. Selección de centreoides para algortimos de clustering a través de técnicas metaheurísticas. In *Congreso Argentino de Ciencias de la Computación*, 2007.

[6] Freitas A.A. A survey of evolutionary algorithms for data mining and knowledge discovery. In *Advances in Evolutionary Computation*. Springer-Verlag, 2001.

[7] Freitas A.A. *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer-Verlag, August 2002.

[8] Elkan C. Using triangle inequality to accelerate k-means. In *ICML*, pages 147–153, 2003.

[9] Pandolfi D, Lasso M., de San Pedro M., Villagra A., and Gallard R. Knowledge insertion: an efficient approach to reduce search effort in evolutionary scheduling. *Journal of Computer Science and Technology*, 4(2):109–114, 2004.

[10] Pelleg D. and Moore A. Accelerating exact k-means algorithms with geometric reasoning. In *Knowledge Discovery and Data Mining*, pages 277–281, 1999.

[11] de San Pedro M., Pandolfi D., Villagra A., and Lasso M. Adaptación dinámica de parámetros en mcmp-sri para el problema de máquina única de weighted tardiness. In *Congreso Argentino de Ciencias de la Computación*, 2006.

[12] de San Pedro M., Pandolfi D., Lasso M., and Villagra A. Dynamic scheduling approaches to solve single machine problem. In *International Conference on Artificial Intelligence and Soft Computing*, 2005.

[13] Lloyd S. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[14] Kanungo T., Nathan S., and Wu A.Y. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):881–892, July 2002.

[15] Zhenjie Zhang, Bing Tian Dai, and Tung A. K. H. On the lower bound of local optimums in k-means algorithm. *IEEE CNF*, pages 775–786, 2006.