Diseño de Algoritmos Evolutivos Híbridos Optimizados para Biclustering

LÍNEA DE INVESTIGACIÓN

Macarena Anahí Latini, Dra. Rocío Cecchini, Dra. Jessica Andrea Carballido

[1] Instituto de Ciencias e Ingeniería de la Computación Universidad Nacional del Sur - CONICET Bahía Blanca - Argentina

E-mail: jac@cs.uns.edu.ar

RESUMEN

El objetivo general de esta línea de investigación consiste en diseñar nuevas técnicas computacionales que ayuden a descubrir potenciales conexiones entre datos presentados en forma de matriz pertenecientes a distintos campos de aplicación. Más específicamente, se planea desarrollar una estrategia evolutiva hibridada con búsqueda local especialmente diseñada para bilcustering de datos. En tal sentido, se busca desarrollar una herramienta que pueda asistir a investigadores de distintas disciplinas en la inferencia de relaciones entre datos procedentes de grandes volúmenes de información.

CONTEXTO

Esta línea se enmarca en el siguiente proyecto de investigación subsidiado por la UNS:

PGI-UNS Tema: Modelado predictivo en Bioinformática basado en el desarrollo de técnicas de Computación Evolutiva y Aprendizaje Automático. Código 24/N042. Entidad financiadora: Secretaría de Ciencia y [illegible] 31/12/2019.

El mismo está dirigido por la Dra. Carballido y se encuentra acreditado para el programa de Incentivos.

Además, la línea también es financiada por el siguiente proyecto:

PIP 2013-2015. Tema: Diseño de Modelos Predictivos en Bioinformática basados en técnicas de Minería de Datos y Aprendizaje Automático. Código: 112-2012-0100471CO. Director: Dr. Ignacio Ponzoni. Entidad financiadora: CONICET. Institución de ejecución: Laboratorio de Investigación y Desarrollo en Computación Científica, UNS. Monto otorgado: $360.000 (trescientossesenta mil pesos). Período de ejecución: Iniciado el 25/8/2014, duración: 3 años.

1. Introducción

Más allá de esto, el biclustering es una metodología que tiene gran variedad de posibles aplicaciones, por lo que se espera abordar en esta línea de investigación un aporte también para alguna de ellas. Por nombrar algunas:

- Marketing: búsqueda de grupos de clientes con un comportamiento similar dado una gran base de datos de datos de clientes que contienen sus propiedades y los registros de compra pasados;
- Seguros: identificación de los grupos de usuarios de seguros con un costo promedio de alta demanda; [illegible] udes;
- Planificación urbanística: identificación de grupos de viviendas de acuerdo al tipo de casa, el valor y la ubicación geográfica;
- Estudios de terremotos: observación de agrupaciones de epicentros sísmicos para identificar zonas peligrosas;
- WWW: clasificación de documentos; la agrupación de datos para descubrir grupos de patrones de acceso similares.

Por último, como objetivo a largo plazo se busca contribuir a la formación de RRHH en Bioinformática, la cual constituye un área científica de vacancia en la Argentina. Asimismo, cabe destacar que esta área está siendo fuertemente impulsada por el LIDeCC[1] a través de actividades de cooperación internacional y ejecución de proyectos bilaterales, y participando en la creación de la Asociación Argentina de Bioinformática y Biología Computacional.

2. Líneas de investigación y desarrollo

Biclustering

Las características comunes principales de las técnicas de clustering se resumen en la búsqueda de conjuntos disjuntos de datos, de tal manera que aquellos datos que se encuentren en un mismo clúster presenten un comportamiento similar frente a todas las columnas de la matriz. Además, cada uno de los datos debe pertenecer a un único clúster (y no a ninguno) al final del proceso. Las técnicas de biclustering [a, b] se presentan como una alternativa más flexible, ya que permiten que las agrupaciones se formen no solo en base a una dimensión, sino que sea posible formar biclusters que contengan datos que presenten un comportamiento similar frente a un subconjunto de las columnas de la matriz. Esta característica es muy importante, ya que aumenta la capacidad de extracción de información a partir de un mismo conjunto de datos, pudiendo ignorar determinadas columnas frente a las cuales un grupo de datos no presenten un comportamiento coherente.

Otro aspecto significativo que diferencia a las técnicas de biclustering frente a las declustering es la forma en que las agrupaciones son hechas, ya que ahora se permite el solapamiento (datos que pueden estar contenidos en varios biclusters a la vez), así como que existan datos que no se hayan incluido en ningún subconjunto. Esta característica aporta más flexibilidad a este tipo de técnicas, ya que no obliga a incluir cada dato en una agrupación determinada, sino que un determinado dato no pertenecerá a ningún bicluster si su valor no se ajusta a ninguno de los patrones.

Un punto importante consiste en determinar la medida de evaluación que permite analizar la calidad de los biclusters. Existen medidas generales, y se pueden definir medidas ad-hoc de acuerdo al problema particular siendo resuelto. En este sentido, la Dra. Carballido realizó recientemente una revisión teórica de métodos evolutivos para biclustering y las distintas medidas de evaluación utilizadas, la cual fue presentada en [c]. Allí se sugiere como medida de evaluación más completa el Error Virtual [d], ya que puede detectar tanto biclusters que presenten patrones de desplazamiento como de escalado. Otro tipo interesante de bilcusters son los biclusters basados en una evolución coherente de sus valores. Algunos algoritmos de biclustering intentan solucionar el problema de encontrar evoluciones coherentes a lo largo de las filas o columnas de una matriz sin tener en cuenta sus valores exactos. Esta forma totalmente distinta de considerar a un bicluster se tiene en cuenta en importantes algoritmos [e, f, g]. Estos serán analizados en una de las aplicaciones propuestas en este plan, donde se buscan perfiles compartidos entre filas de la matriz (que representan expresión de genes, como se verá más adelante).

Algoritmos genéticos: hibridación y optimización de operadores

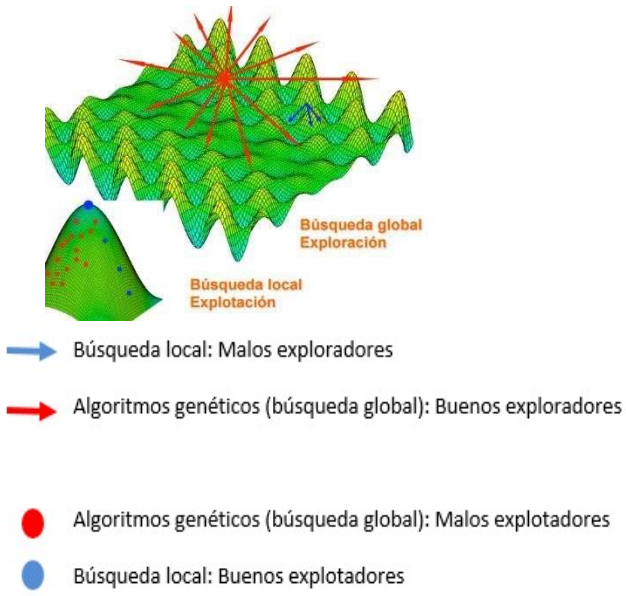
Un algoritmo memético [h] es la combinación entre:

- Una búsqueda global basada en poblaciones,
- Una heurística de búsqueda local realizada por cada individuo.

Es importante aclarar que la búsqueda global no implica necesariamente un algoritmo genético. Sin embargo, en este plan se proyecta utilizar esta técnica como base de la metodología desarrollada. La clave está en que los algoritmos genéticos son buenos "exploradores de soluciones" pero malos "explotadores", en cambio los algoritmos de búsqueda local son

[1] Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC), Depto. Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur. Sitio web oficial: http://lidecc.cs.uns.edu.ar

buenos "explotadores" de soluciones prometedoras, mientras que son malos "exploradores" [i]. De esta manera surge la idea de combinarlos. Gráficamente:



Luego, como primer paso, pretendemos profundizar en tres puntos de interés que se deberían tener en cuenta al momento de desarrollar un algoritmo memético a partir de un algoritmo genético. Cómo hibridarlo es una cuestión que se puede responder parcialmente, debido a que es sabido que la hibridación consiste en combinarlo con una estrategia de búsqueda local. Cuál es el método de búsqueda local que en términos generales mejor desempeño produce, será analizado inicialmente. Proponemos también investigar dónde hibridarlo, lo cual será estudiado de acuerdo al mejor momento del proceso de evolución en el cual debería intervenir la búsqueda local. Por último en este sentido, planeamos también responder a la pregunta de cuándo hibridarlo, y con esto nos referimos a los tipos de problemas en los que es aconsejable utilizar este tipo de estrategia sinérgica.

Por otro lado, en lo que refiere a la mejora en la calidad de los operadores, se proyecta estudiar una forma de cruzamiento inteligente que aproveche el conocimiento de las características de las técnicas de biclustering en el proceso de reproducción. Para esto planteamos la hipótesis de realizar un proceso recursivo que permita combinar de manera eficaz dos o más biclusters produciendo uno de mejor calidad que sus ancestros [j]. Una de las ideas surge de observar la efectividad del método de Ward implementado en el algoritmo recursivo de Lance-Williams [k]. La recursividad aplicada al biclustering se fundamenta en que esta técnica es ideal para objetos estáticos ya que se puede calcular previamente la medida de calidad para distintos grupos y utilizar estas medidas para instancias posteriores. Asimismo, se estudiará en detalle el criterio de terminación del algoritmo, para asegurar la convergencia a soluciones variadas en el contexto de cada problema [l].

Aplicación en Bioinformática: Datos de expresión de genes

La tecnología de microarray permite analizar niveles de expresión de miles de genes en distintas condiciones. Esta información se presenta en forma de matriz de números reales donde cada elemento representa el nivel de expresión del gen en la condición correspondiente [m]. Una de las aplicaciones más populares de las técnicas de biclustering está justamente relacionada a la extracción de información de genes que se expresan de manera coherente en datos de microarray [m]. Un ejemplo particular de esto es la inferencia de redes de regulación de genes (GRNs) usando biclustering, que representan relaciones entre genes que podrían constituir agrupamientos encargados de alguna función celular determinada [n]. En este sentido, la directora propuesta viene participando en el último período de investigaciones orientadas a dicho tema de investigación [o, p]. Luego, en el contexto de este plan, otra de las hipótesis planteada es la de encontrar pares de genes co-expresados; es decir, que tienen un perfil de expresión similar. De esta forma, para una aplicación de *eSalud* muy promisoria, se podría usar la metodología para inferir nuevos posibles marcadores tumorales a partir de marcadores conocidos [q, r].

3. Resultados esperados

Con esta línea de investigación se espera principalmente lograr una contribución teórica en algoritmos meméticos, a partir del diseño y desarrollo de una nueva estrategia híbrida que combine de manera eficaz las mejores

características de los algoritmos evolutivos y de un método de búsqueda local, mejorando así las capacidades de búsqueda global y local de cada una de estas estrategias respectivamente. Se diseñará la metodología con el fin de encontrar uno o varios biclusters en grandes cantidades de datos contenidos en una matriz. En principio, en la búsqueda local se analiza la posibilidad de incorporar un proceso recursivo para la reparación y/u optimización de los biclusters encontrados.

Cabe destacar la experiencia de la Dra. Carballido en el área de computación evolutiva, temática que ha desarrollado desde su propia tesis doctoral hasta la actualidad, aplicándola a distintos problemas de optimización. También ha tenido un primer acercamiento a la hibridación básica de algoritmos genéticos en una beca doctoral dirigida con anterioridad.

Asimismo, la Dra. Carballido ha dedicado los últimos años de su trabajo al estudio de técnicas de biclustering para datos de expresión obtenidos de experimentos de microarray. Por este motivo, una de las aplicaciones a abordar en una primera instancia pertenece al área de bioinformática. En particular, el problema que se espera atacar consiste en encontrar relaciones entre distintos genes a partir de información sobre su nivel de actividad. El desarrollo de metodologías con estas características podría resultar de suma utilidad en la investigación del área de *eSalud*. En este sentido ya está comprobada la eficacia de los enfoques de biclustering para seleccionar conjuntos de genes óptimos para la determinación de pronósticos de estratos específicos de pacientes en base a características moleculares de tumores. En particular, para este plan de trabajo se plantea la siguiente hipótesis: dado un gen conocido que se sabe que constituye un marcador tumoral comprobado, otro gen con un perfil de expresión similar podría ser también asociadoa dicha enfermedad. El perfil de expresión similar será encontrado con la técnica de biclustering.

4. Formación de recursos humanos

En el contexto de esta línea intervienen el Dr. Ignacio Ponzoni (Investigador CONICET), la Dra. JessicaCarballido (Investigadora CONICET), la Dra. Rocío Cecchin (Investigadora CONICET), la Dra. Julieta Dussaut (becaria POSDOC CONICET) y la Ing. Jimena Martínez (becaria Doctoral CONICET) principalmente. Todos sus planes de investigación están estrechamente vinculados a esta línea. Además constituye el eje principal del plan de doctorado de la Ing. Macarena Latini.

De este modo, esta línea se suma contribuirá a la formación de integrantes en distintosniveles de CONICET, en una disciplina de fuerte proyección en la actualidad, tal como es bioinformática. Con lo cual, nuestro laboratorio (LIDeCC) seguirá afianzándose en la formación de nuevos investigadores, tal como ha sido su tradición desde su conformación como grupo de investigación en el año 1996.

5. Referencias

a. Madeira SC, Oliveira AL (2004). "Biclustering Algorithms for Biological Data Analysis: A Survey". IEEE Transactions on Computational Biology and Bioinformatics 1 (1): 24–45. doi:10.1109/TCBB.2004.2. PMID 17048406.

b. Kriegel, H.-P.; Kröger, P.; Zimek, A. (March 2009). "Clustering High Dimensional Data: A Survey on Subspace Clustering, Pattern-based Clustering, and Correlation Clustering". ACM Transactions on Knowledge Discovery from Data 3 (1): 1–58. doi:10.1145/1497577.1497578.

c. Carballido J.A., Gallo C.A., Dussaut J.S., Ponzoni I. “On Evolutionary Algorithms for Biclustering of Gene Expression Data”, Current Bioinformatics (2015). 10 3 259-267(9).

d. Beatriz Pontes, Federico Divina, Raúl Giráldez, Jesús S. Aguilar–Ruiz (2007), Virtual Error: A New Measure for Evolutionary Biclustering, Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics Vol 4447 of the series Lecture Notes in Computer Science pp 217-226

e. Everitt, B. S., Landau, S. and Leese, M. (2001), Cluster Analysis, 4th Edition, Oxford University Press, Inc., New York; Arnold, London.

f. Hartigan, J. A. (1975), Clustering Algorithms, New York: Wiley.

g. Jain, A. K. and Dubes, R. C. (1988), Algorithms for Clustering Data, New Jersey: Prentice–Hall.

h. Moscato, P., Cotta, C.: A gentle introduction to memetic algorithms. In Glover, F., Kochenberger, G., eds.: Handbook of Metaheuristics. Kluwer Academic Publishers, Boston MA (2003) 105–144.

i. Ishibuchi, H., Yoshida, T., Murata, T.: Balance between genetic search and local search in memetic algorithms for multiobjective permutation flowshop scheduling. IEEE Transactions on Evolutionary Computation 7 (2003) 204–223.

j. Eiben, A.E., Raue, P.E., Ruttkay, Z.: Genetic algorithms with multi-parent recombination. In Davidor, Y., Schwefel, H.P., M¨anner, R., eds.: Parallel Problem Solving From Nature III. Volume 866 of Lecture Notes in Computer Science. Springer-Verlag (1994) 78–87

k. Cormack, R. M. (1971), "A Review of Classification", Journal of the Royal Statistical Society, Series A, 134(3), 321-367.

l. Safe M.D., Carballido J.A., Ponzoni I, Brignole N.B. “On Stopping Criteria for Genetic Algorithms” Lecture Notes in Artificial Intelligence, Vol. 3171, 405–413 (2004). Springer-Verlag.

m. Kluger Y., Basri R., Chang J., and Gerstein M. (2003). Spectral biclustering of microarray data: co-clustering genes and conditions. Genome Research, 13:703–716.

n. Ma, S.; Kosorok, M.R. “Identification of differential gene pathways with principal component analysis”, Bioinformatics, 25:882-889, 2009.

o. Gallo C.A., Cecchini R.L., Carballido J.A., Micheletto S., Ponzoni I. “Discretization of gene expression data revised”. Briefings in Bioinformatics (2015). 1-13.

p. Gallo C.A., Carballido J.A., Ponzoni I. “Discovering Time-Lagged Rules from Microarray Data using Gene Profile Classifiers”, BMC Bioinformatics (2011). 12:123.

q. Wang YK, Print CG, Crampin EJ (2013) Biclustering reveals breast cancer tumour subgroups with common clinical features and improves prediction of disease recurrence. BMC genomics 14: 102. Ali Oghabian, Sami Kilpinen, Sampsa Hautaniemi, Elena Czeizle, Biclustering Methods: Biological Relevance and Application in Gene Expression Analysis http://dx.doi.org/10.1371/journal.pone.0090801