Minería de Datos con Redes Neuronales Artificiales: Aplicación en Vacunas – Tuberculosis.

M.V. Guzmán (*), H. Carrillo (**), E. Villaseñor (**), E. Valencia (**), R. Calero (*), L. E. Morán (**) y A. Acosta (*).

* Instituto Finlay. Centro de Investigación-Desarrollo y Producción de Vacunas y Sueros. Ave. 27 No. 19805, La Lisa. La Habana. A.P. 16017 Cod. 11600. Telf: 331218 / email: mvguzman@finlay.edu.cu.
** Laboratorio de Dinámica no Lineal, Facultad de Ciencias, UNAM, México. email: carr@servidor.unam.mx

El volumen de datos que se acumula continuamente, y la necesidad de encontrar métodos que permitan descubrir conocimiento (dentro de esas enormes masas de datos), han convertido a la Minería de Datos en una disciplina de importancia estratégica para la planeación y la toma de decisiones. La Minería de Datos se apoya en la aplicación de métodos matemáticos de análisis, y específicamente del uso redes neuronales artificiales, que son de gran utilidad para llevar a cabo el análisis inteligente de grandes volúmenes de información digital. En este trabajo se revisan y discuten algunos aspectos de esta metodología y se ilustra su aplicación investigando un volumen considerable de textos digitales (2987 artículos de las bases de MedLine) referentes a la investigación de vacunas contra la tuberculosis.

Temáticas: *Minería de datos, Minería de textos, Análisis inteligente de datos, Redes neuronales artificiales, Bases de datos, Vacunas tuberculosis.*

Introducción

La tecnología moderna permite la creación de grandes almacenes de datos (crudos) que requieren ser explorados en búsqueda de información refinada (conocimiento). Desarrollar agentes que permitan procesar estos grandes volúmenes de datos y convertirlos en conocimiento útil para la toma de decisiones (inteligencia), constituye un reto colosal. Nuevas disciplinas han emergido para abordar este problema: Descubrimiento de Conocimiento (Knowledge Discovery), Minería de Datos (Data Mining), Análisis Inteligente de Datos (Inteligent Data Análisis), Análisis Exploratorio de Datos (Exploratory Data Análisis) [1], [2], [3], [4]. Estas disciplinas se basan en métodos de la Matemática y de la Inteligencia Artificial para acometer esta nueva problemática. Las tecnologías desarrolladas para el procesamiento de la información han tenido un impacto revolucionario en la industria y en el mundo de los negocios. Hoy en día existe una gran variedad de sistemas

de software comerciales que se basan en las técnicas del Análisis Inteligente de Datos para llevar a cabo tareas como: planeación económica, vigilancia e inteligencia empresarial, análisis financiero, análisis de mercados y análisis de perfiles de clientes.

Esta revolución de la Tecnología de la Información coincide con la revolución que paralelamente está teniendo lugar en la Biotecnología, debido a un mundo de avances tecnológicos que han permitido la obtención y acumulación automática de inmensas cantidades de información biológica. Del encuentro de estas dos revoluciones nace la Bioinformática como una multidisciplina estratégica que pone al servicio de la Biotecnología los recursos de la nuevas Tecnologías de la Información. Con todo este arsenal (de datos, métodos y software) hoy se hacen sofisticadas investigaciones para organizar la masa de información disponible con el fin de conceptualizar y fundamentar la biología en términos de los principios de la físico-química molecular. Este afán confronta grandes dificultades y requiere el concurso de diferentes herramientas entre los cuales el análisis matemático juega un papel fundamental [5],[6],[7]. La computación y visualización que permiten estas herramientas se usa para analizar secuencias genómicas, estructuras macromoleculares y datos de expresión.

La Bioinformática toma datos de diversas fuentes una de las más utilizadas son las provenientes del National Center for Biotechnology Information (NCBI). Esta institución facilita (online) información del proyecto genoma (GenBank®), bases de datos de sobre biología molecular (Nucleotides, Proteins, Structures, Genes expression, Taxonomy), sobre literatura biomédica (PubMed, PubMedCentral, OMIM, Books, Citation Matcher), etc.

Figura 1. Ejemplo de base de datos utilizadas para estos fines como Taxonomy Browser.

Las nuevas tecnologías de manejo de información y particularmente los recursos del Análisis Inteligente de Datos, son también de gran utilidad para el análisis de información científica textual (documentos digitales), como la que se tiene en las bases de datos de patentes o artículos de investigación (e.g.: MedLine o Biological Abstracts). En estas investigaciones de "Minería de Textos" convergen también los métodos de la Cienciometría, la Bibliometría y la Infométría, para aportar valiosos conceptos, indicadores y técnicas de análisis [8], [9] y [10].

Procesos y Métodos

El descubrimiento de conocimiento en bases de datos de información científica puede ser entendido como un proceso que implica la realización de una secuencia básica de tareas:

- Comprensión del campo de aplicación
- Adquisición y selección de ficheros.
- Preprocesamiento de ficheros.
- Minería de Datos de resultados.
- Visualización e interpretación de resultados.
- Evaluación y reporte de resultados.

Este proceso de descubrimiento de conocimiento es iterativo e interactivo y tiene a la minería de datos como una de sus principales etapas [11]. La minería de datos integra métodos estadísticos con métodos de "aprendizaje maquina", y en particular redes neuronales, para llevar a cabo el proceso de análisis exploratorio de datos [12]. Consultar el conocimiento y la experiencia de los expertos es una fase importante del ciclo de análisis. Estos deben participar interactivamente en el proceso de interpretación, visualización, evaluación y reporte de resultados.

Inspirados en la anatomía y fisiología del cerebro humano, las Redes Neuronales Artificiales (RNA) son modelos matemáticos que permiten hacer computación inteligente [13] y llevar a cabo tareas que las computadoras seriales no pueden realizar: reconocimiento de patrones, memorias y aprendizaje asociativo, control adaptivo, predicción de series de tiempo, clasificación de señales y clustering, entre otras.

En una computadora neuronal el procesamiento es distribuido a toda una red de procesadores denominados "neuronas" que realizan el cómputo en paralelo. Las propiedad de distribución y la capacidad de paralelizar los procesos determinan las nuevas capacidades implicadas en el paradigma neuronal. Desde el punto de vista de la minería de datos [14], el procesamiento paralelo

y distribuido es muy importante porque permite que las redes neuronales sean capaces de llevar a cabo el procesamiento de datos a una escala masiva.

El SOM (Self-Organizing Map) es un eficiente algoritmo neuronal (no supervisado) que permite proyección de datos que habitan en un espacio multidimensional, a una retícula bidimensional denominada “mapa”, preservando cualitativamente la organización (topología) del conjunto original. Desde que el SOM fue introducido por T. Kohonen en el año 1982 [15], sobre este algoritmo se han producido una gran cantidad de artículos de investigación [16], y basados en él, se han desarrollado diversas aplicaciones de software para la minería de datos [17], que han sido aplicadas a la solución de una gran variedad de problemas [17].

Un ejemplo notable dentro del conjunto de herramientas de software que usan el SOM, es el sistema Viscovery SOMine, de la compañía austriaca, Eudaptics Sofware Company [18]. Este sistema tiene una interfaz amigable e interactiva con el usuario, que facilita la generación automática de “mapas de conocimiento”, como veremos a continuación.

Aplicaciones

Hace unos cinco años, un grupo de investigadores del Laboratorio de Dinámica no Lineal de la Universidad Nacional Autónoma de México y del Instituto Finlay de Cuba, nos hemos avocado a explorar las bases digitales de información biomédica aprovechando las nuevas tecnologías para el análisis inteligente de datos y el descubrimiento de conocimiento, salvando así las limitaciones que tiene la aplicación de otros métodos tradicionales a tan grandes volúmenes de datos. Durante este lapso se ha ido desarrollando toda una metodología que ocupa diversos sistemas de software y se ha ido aplicando experimentalmente en la realización de diversas investigaciones. En lo que sigue, a modo de ejemplo, revisaremos un par de estas aplicaciones.

A. Análisis de la relevancia de diferentes sustancias químicas en las investigaciones sobre la tuberculosis.

La investigación de vacunas contra la tuberculosis se ha vuelto un problema de gran actualidad ya que se trata de una enfermedad re-emergente para la cual no se cuenta aún con alguna vacuna suficientemente efectiva. De acuerdo a la Organización Mundial de la Salud (OPS) esto constituye una situación de emergencia para nuestra orbe.

A los especialistas, que trabajan en vacunas contra la tuberculosis, les interesa conocer la forma en que está evolucionando el uso de diferentes sustancias en este tipo de investigaciones, a nivel mundial. Haciendo uso de las técnicas de la Minería de datos nuestro grupo analizó 2987 artículos de investigación contenidos en las bases de datos de MedLine (literatura biomédica) e investigó el uso de 8,961 diferentes sustancias que aparecen reportadas en las investigaciones de un lapso de 22 años (1980-2002).

Se observó que no todas tenían la misma incidencia en los 22 años de análisis y el estudio de frecuencias de ocurrencia reveló que, en la década de los 80´s las sustancias en las que se trabajaba más eran los agentes antineoplásticos (45 investigaciones) y la Ciclofosfamida (32 investigaciones), mientras que en la década de los años 90´s otras sustancias, como los adyuvantes inmunológicos, los interferones y los antígenos pasaron a ocupar los primeros lugares (208, 116 y 106 respectivamente). Asimismo se concluyó que actualmente (período 2000-2002), los adyuvantes (128) y los interferones (102) se siguen utilizando pero también se observa una emergente tendencia a la investigación de vacunas sintéticas y de DNA.

Dado el interés de los investigadores asociados con nuestro grupo, posteriormente especializamos la investigación a una familia de sustancias que tienen efecto en la modificación de la respuesta inmuno-biológica: las interleucinas (Interleukins). El propósito de este estudio fue, primero, identificar los diferentes tipos de Interleucinas, que son considerados en las investigaciones sobre vacunas contra la tuberculosis y después estudiar la evolución que ha tenido su utilización durante el período de análisis (1980-2002). Se identificaron las sustancias Interleukin-1, Interleukin-2, Interleukin-4, Interleukin-6 y Interleukin-12 en un conjunto de 2,600 sustancias (1600 resultados de investigación).

Entrenando una red neuronal (usando el sistema de software Viscovery SOMine) se generó mapas específicos para representar las sustancias relacionadas con la Interleukina-1 y la Interleukina-12. A continuación (Figura 1 y 2) se desplegaron los mapas correspondientes a los períodos (1990-1999) y (2000-2002). A pesar de que la interleukina-1 apareció con una frecuencia mayor que la Interleukina-12, en el período 2000-2002, los mapas producidos (Figura 1) exhiben claramente que esta última sustancias aparece asociada a un número considerablemente mayor de sustancias.

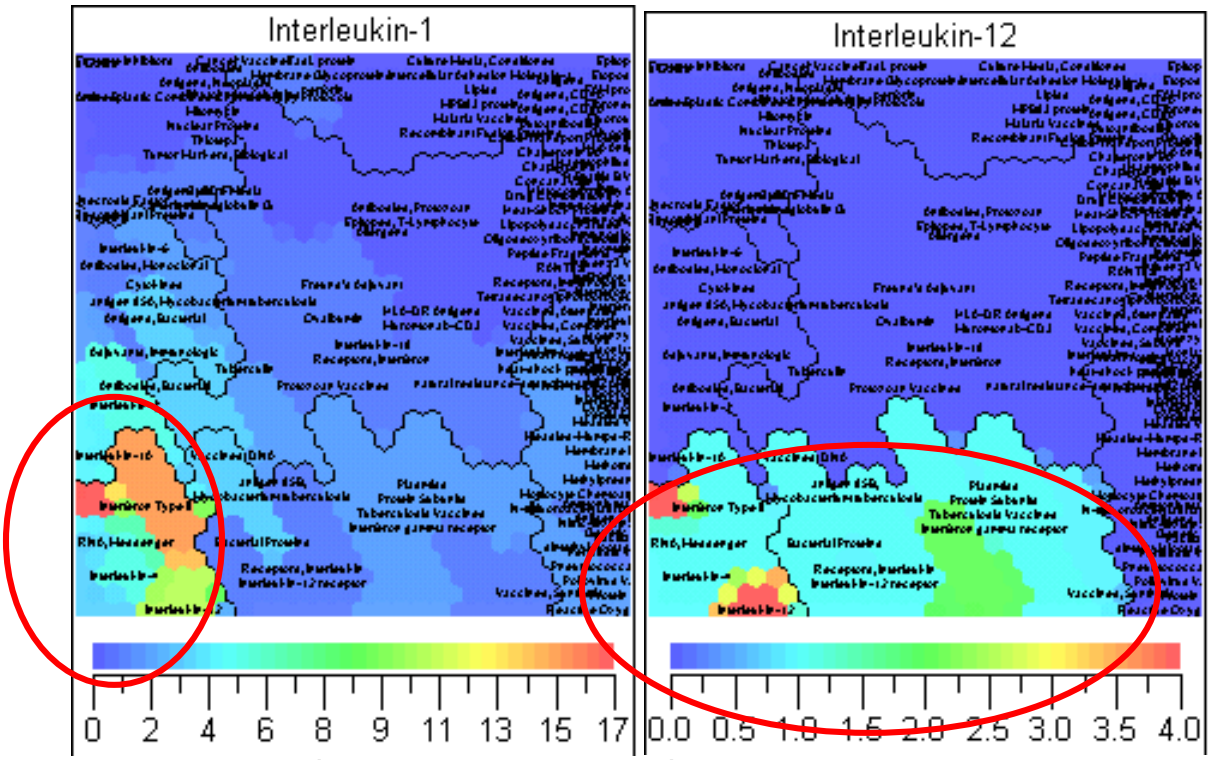

Figura 2. Representación del uso en la investigación de las sustancias Interleukin-1 e Interleukin-12 para el periodo 2000-2002.

Por otra parte, el análisis retrospectivo del lapso (1990-1999) mostró que la aparición de la Interleukina-1 predominaba respecto a la Interleukina-12, dado que esta última se asociaba con muy pocas sustancias en este período. Compárese los resultados de la Figura 1 con los obtenidos en la Figura 2. (2000-2002).

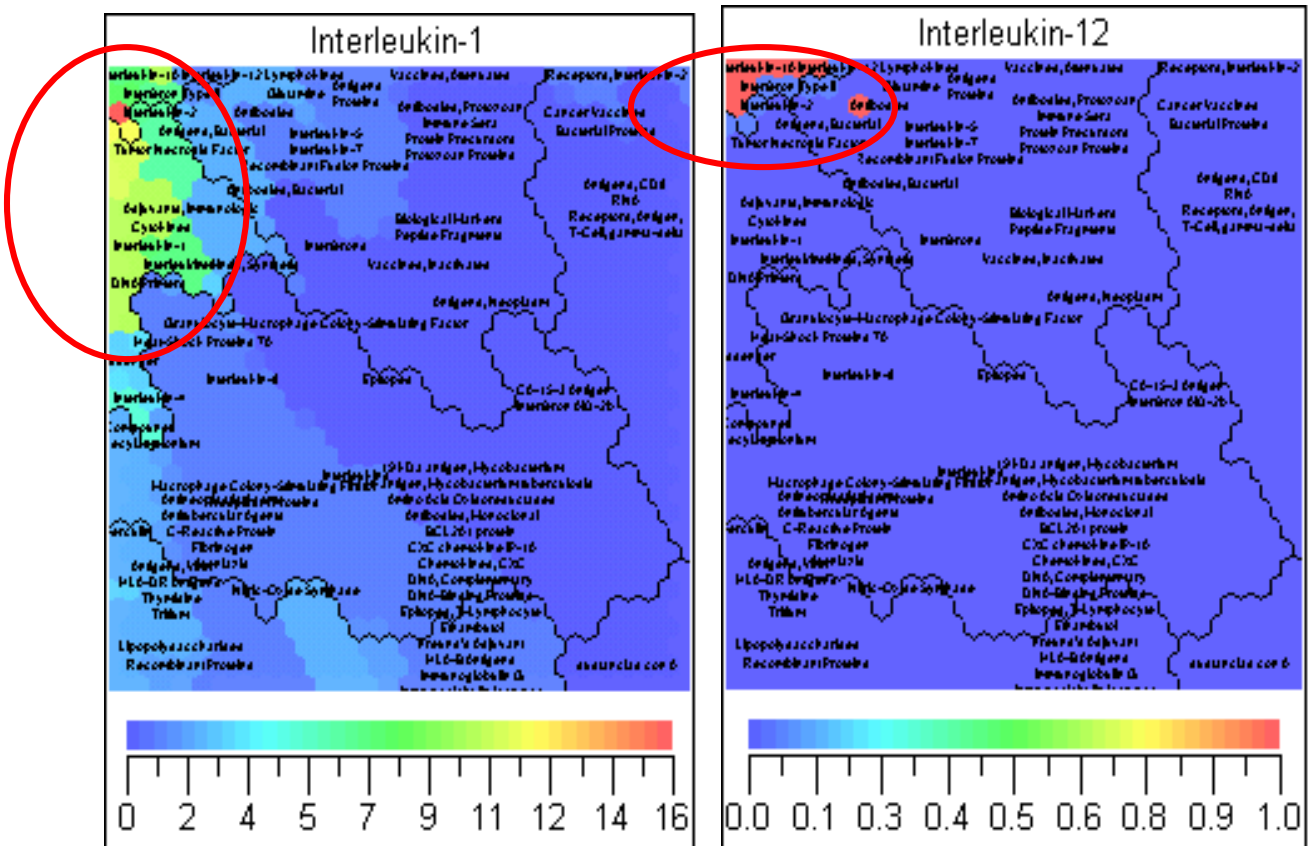

Figura 3. Representación del uso en la investigación de las sustancias Interleukin-1 e Interleukin-12 para el periodo 1990-1999.

Lo mostrado son solo dos ejemplos de las análisis que se pueden hacer basado en el principio de la Minería de datos y textos. Estos son validos para otros campos del conocimientos, solo se necesita identificar el problema y aplicar el modelo correspondiente.

El desarrollo las técnicas de minería de datos y textos asociadas a la bioinformática se han convertido en ejes de desarrollo de la Biotecnología sin embargo calcular su impacto económico resulta difícil. A pesar de ello y ante el hecho evidente del largo lapsos de tiempo que demora un producto biotecnológico en salir al mercado, es evidente que estas técnicas le permitirían agregarle un valor al producto biotecnológico, reducir los tiempos de respuestas a un problema de salud determinado o investigativo pero sobre todo puede marcar la diferencia en un modo de hacer ciencia.

Bibliografía

[1]. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From data mining to knowledge discovery: an overview", in Fayyad, U., Piatetsky-Shapiro, G., Smyth, P and Uthurusamy, R. (Eds), Advances in knowledge Discovery and Data Mining, MIT Press, Cambridge, M.A. 1996

[2]. M.J. Norton, "Knowledge Discovery in Database", Library Trends, 48(1):9-21, 1999.

[3]. J. W. Tukey, "Exploratory data analisys", Addison Wesley, 1977.

[4]. M. Berthold, D.J. Hand, Intelligent data analysis, Springer, 2000

[5]. H. Carrillo, D. Lípman, "The Multiple Sequence Alignment Problem in Biology", SIAM Journal of Applied Mathematics, Vol.48, No. 5, pp. 1073-1082, 1988. ().

[6]. K. Reinert, J. Stoye, T. Will, "An Iterative Method for Faster Sum-of-pairs multiple sequence alignment", Bioinformatics Vol. 16, No. 9, pp. 808-814, 2000.

[7]. J. Kececioglu, "Notes on a Múltiple Alignment cost-bound of Carrillo and Lipman", Proceedings of the 6th Symposium on Combinatorial Pattern Matching, Springer Verlag, Lecture Notes on Computer Science 937, 128-143, 1995

[8]. G. Sotolongo, M. V. Guzmán, "Aplicaciones de las redes neuronales. El caso de la bibliometría", Ciencias de la Información. 2001; 32(1):27-34.

[9]. G. Sotolongo, M. V. Guzmán, H. Carrillo, "ViBlioSOM: visualización de información bibliométrica mediante el mapeo autoorganizado", Revista Española de Documentación Científica, 2002, 25(4):477-484.

[10]. G. Sotolongo, M. V. Guzmán, O. Saavedra, H. Carrillo, "Mining Informetrics Data with Self-organizing Maps", in: M. Davis, C.S. Wilson, (Eds.), "Proceedings of the 8 th International Society for for Scientometrics and Informetrics", ISBN:0-7334-18201. Sydney, Australia July 16-20. Sydney: BIRG; 2001: 665-673.

[11]. J. L. Sang, K. Siau, "A Review of data mining techniques", MCB University Press, p.p. 41-46, 2001.

[12]. H. Mannila, "Data Mining: Machine Learning, Statistics and Databases", Dep. of Computer Science, University of Helsinky.

[13]. A. K. Jain, J. Mao, K. Mohiuddin, "Artificial Neural Networks: A Tutorial", IEEE Computer Special Issue on Neural Computing, 1996.

[14]. J. Bigus, "Data Mining with neural networks", Mc GrawHill, USA, 1996

[15]. T. Kohonen, "Self-organized formation of topologically correct feature maps", Biological Cybernetics, 43, 1982.

[16]. Listado de publicaciones puede ser consultado en: http://www.cis.hut.fi/nnrc/refs/

[17]. T. Kohonen, "Self-Organizing Maps", 3ra Edición, Springer-Verlag, 2001.

[18]. Visitar http://www.eudaptics.com