DATA MINING ŞI CERCETAREA ŞTIINȚIFICĂ

Dr. ing. Cristina - Maria DABU
Inginer programator

A absolvit Facultatea de Automatică şi Calculatoare din Institutul Politehnic Bucureşti în 1992. A obținut titlul de doctor în inginerie industrială la Universitatea Politehnica din Bucureşti, în anul 2001. Are peste 30 lucrări publicate în țară şi în străinătate.

REZUMAT. Cercetările ştiinţifice din ultimii ani în domeniul biologiei moleculare şi medicinei moleculare au generat o cantitate enormă de date cum ar fi secvenţele genomice rezultate din Proiectul Genomul Uman, expresii ale genelor din experimentele cu micromatrici, identificarea proteinelor şi cuantificarea datelor din experimente proteomice, date privind polimorfismele mononucleotidice din matricile de polimorfisme mononucleotidice cu rată înaltă de transfer. Tehnicile de identificare a cunoştinţelor devin din ce în ce mai importante odată cu creştera cantităţii de date colectate. Cantitatea mare de date rezultate din studiile din domeniul viului implică necesitatea dezvoltării de instrumente bioinformatice complexe şi tehnici data mining pentru o analiză eficientă şi precisă a rezultatelor experimentale Lucrarea prezintă ultimile rezultate în domeniul bioinformaticii şi al tehnicilor data-mining precum şi limitările acestora în aplicarea asupra seturilor de date în scopul evidenţierii strânsei interdependente între tehnologiile de cercetare şi instrumentele bioinformatice.

Cuvinte cheie: data mining, bioinformatică, modele, biodate, algoritmi, clustering, rețele neuronale.

ABSTRACT. The scientifically research from the last years in the fields of molecular biology and molecular medicine has generated enormous amounts of data like genomic sequences resulted from the Human Gemome Project, gene expresions from microarray experiments, protein identification and data quantification form proteomics experiments, SNP data from high-throughput SNP arrays. Knowledge-discovery techiques are becoming more and more important as the collected data increased. The extended studies in the fields of proteins and genetics both of qualitative and quantitative identification and comaprison. The huge dimension of data generated from these studies requires the development of improved bioinformatics tools and data-mining approaches for efficient and accurate data analysis. There is a strong interest in employng methods of knowledge discovery and data mining to generate models of biological systems. This paper will present the recent developments in bioinformatics and data-mining approaches and their limitations when applied to proteomics data sets, in order to strengthen the interdependence between proteomic technologies and bioinformatic tools.

Key-words: data mining, **bioinformatics**, models, biodata, algorithms, clustering, neuroal networks.

1. INTRODUCERE

Complexitatea considerabilă ce caracterizează sistemele vii determină necesitatea unei cantități uriaşe de informație detaliată pentru descrierea lor completă. Odată cu progresul tehnologic în domeniul ştiinţelor vieții, cantități uriaşe de date biologice au devenit disponibile sub formă de baze de date publice. Însăşi cheia progresului în cercetare în domeniul ştiinţelor vieții constă tocmai în existența acestor uriaşe baze de date, conținând:

- secvențe genomice rezultate din Proiectul Genomul Uman;
- secvențele de AND/ARN şi secvențele de proteine;
- informații privind căile de semnalizare şi căile metabolice;
- expresii ale genelor din experimente pe bază de micromatrici;
- identificarea proteinelor şi cuantificarea datelor rezultate din experimente proteomice;
- datele privind polimorfismele mononucleotidice rezultate din matrici de polimorfisme mononucleotidice cu rată înaltă de transfer;
- bazele de date cuprinzând literatura biomedicală;
- bazele de date conținând imagini etc.

Principala abordare în analizarea, modelarea şi interpretarea datelor biologice constă în abstractizarea lor în structuri logice care sprijină şi promovează dezvoltarea unui cadru general conceptual pentru caracterizarea, explicarea şi predicția proceselor din cadrul sistemelor vii. Cantitățile enorme de date precum şi necesitatea valorificării lor, au determinat dezvoltarea de aplicații din domeniul inteligenței artificiale, al analizei inteligente a imaginilor, data mining, text mining, reprezentarea şi managementul cunoştinţelor etc., pentru realizarea de operații de genul:

- preprocesarea de task-uri cum ar fi curățarea datelor sau integrarea datelor aplicate datelor biologice;
- tehnici de clasificare şi clustering pentru micromatrice;

- compararea structurilor de RNA pe baza proprietăților şirurilor şi a proprietăților energetice;
- descoperirea secvențelor caracteristice ale diferitelor părți ale genomului;
- analiza haplo-tipurilor pentru identificarea markerilor bolilor;
- secvențierea evenimentelor ce conduc la plierea proteinelor;
- interferența localizării subcelulare a activității proteinelor;
- clasificarea compuşilor chimici în funcție de structură;
- metrici pentru scopuri speciale şi structuri index pentru aplicații filogenetice;
- un nou limbaj de interogare pentru căutarea proteinelor bazat pe forma proteinelor;
- scheme de indexare foarte rapidă pentru secvențe şi căi de comunicare.

Studiile extinse în domeniul proteinelor şi al geneticii de exemplu implică identificări atât din punct de vedere calitativ, cât şi din punct de vedere cantitativ. Volumul mare de date generat de aceste studii implică necesitatea dezvoltării de instrumente bioinformatice puternice, bazate pe tehnici de data-mining pt analiza eficientă riguroasă a seturilor mari de date.

O posibilă definiție a data-mining-ului ar putea fi „extragerea de potențiale informații, necunoscute în prealabil şi posibil utile, din date existente, prin metode nebanale" [1].

2. TEHNICI DE DATA MINING FOLOSITE ÎN CERCETARE

Abordările din Data mining se dovedesc a fi ideale pentru problemele specifice bioinformaticii, un domeniu bogat în date, dar sărac în teorii cuprinzătoare privind organizarea vieții la nivel molecular.

În schimb, cantitățile mari de date şi informații biologice specifie domeniului oferă posbilitatea realizării de noi cercetări şi obținerea de noi metode de analiză a datelor.

Data mining implică dezvoltarea unor instrumente sofisticate pentru: analiza datelor pentru a identifica patternuri valide şi relații necunoscute în prealabil, în cadrul unor seturi mari de date. Aceste instrumente pot include:

– modele statistice;

– algoritmi matematici şi metode ale inteligenței artificiale (machine learning methods);

– algoritmi care îşi îmbunătățesc performanțele în mod automat pe baza experienței (rețelele neurale, arborii de decizii).

Data mining înseamnă mai mult decât colectarea datelor şi gestionarea datelor, ea mai implică analiză şi predicție şi se poate face pe date reprezentate în diverse forme (cantitativ, text, multimedia). În plus, una dintre problemele principale care se ridică în timpul procesului de obținere a datelor şi rămase nerezolvate este tratarea datelor ce conțin informații temporale. În acest caz, o înțelegere deplină a întregului fenomen presupune ca datele să fie privite ca o succesiune de evenimente. Scopul fundamental al data-mining-ului temporal este descoperirea de relații ascunse între secvențe şi subsecvențe de evenimente. Descoperirea relațiilor dintre secvențele de evenimente comportă, îndeosebi, trei paşi:

– reprezentarea şi modelarea secvenței de date într-o formă adecvată;

– definirea măsurilor de similaritate între secvențe;

– aplicarea modelelor şi reprezentărilor problemelor de mining actuale.

În funcție de natura secvenței de evenimente, abordările problemelor pot diferi.

Tehnicile de data-mining pot folosi o varietate de parametri pentru examinarea seturilor de date, incluzând:

– identificarea asocierilor între evenimente;

– analiza secvențelor sau a căilor de transmitere a diverselor semnale (la nivel biologic – un eveniment declanşează producerea unui alt eveniment);

– clasificări - identificări de noi patternuri;

– clustering (identificarea şi documentarea vizuală a unor grupuri de evenimente necunoscute în prealabil)

– prognoză (descoperirea de patternuri din care se pot face predicții rezonabile privind comportamente ulterioare).

Data mining utilizează o abordare prin descoperire în cadrul căreia algoritmii pot fi utilizați pentru a examina simultan mai multe relații multidimensionale între date, identificând pe acelea care sunt unice sau care îşi fac apariția în mod frecvent.

Data mining poate fi considerată doar un pas în marele proces cunoscut ca „descoperirea de cunoştințe în bazele de date" (knowledge discovery în databases – KDD).

Paşii KDD în ordine progresivă sunt [2]:

- curățirea datelor;
- integrarea datelor;
- selecția datelor
- transformarea datelor;
- data mining;
- evaluarea patternurilor
- prezentarea cunoştințelor.

O serie de elemente ce țin de avansul tehnologic au contribuit la creşterea interesului în domeniul data mining atât în sectorul public, cât şi în cel privat. Printre acestea, se numără:

- dezvoltarea rețelelor de calculatoare care pot fi utilizate pentru conectarea bazelor de date;
- dezvoltarea de tehnici avansate de căutare asociate, cum ar fi rețelele neurale şi algoritmii avansați;
- utilizarea pe scară tot mai largă a modelelor client server ce permit utilizatorilor să aibă acces centralizat la sursele de date de la propriul sistem de calcul;
- capacitatea din ce în ce mai mare de a combina date din surse disparate într-o singură sursă ce poate fi gestionată şi asupra căreia se pot efectua operațiuni de căutare;
- creşterea disponibilității informației;
- reducerea costurilor de stocare au jucat un rol important.

Când aceste tehnici sunt implementate pe sisteme de procesare client/server sau paralele de mare performanță , pot analiza baze de date masive, specifice de exemplu bioinformaticii.

Tehnologia data mining se caracterizează prin:

- analiză automată;
- seturi de date largi sau complexe;
- descoperirea de tipare semnificative sau tendințe care altfel ar trece neobservate.

Tehnologia Data Mining este pregatită pentru aplicații deoarece ea este susținută de 3 tehnologii care acum sunt suficient de dezvoltate :

- sisteme de gestiune pentru colecții masive de date;
- computere multi-procesor puternice;
- existența algoritmilor specifici de căutare a datelor.

3. TEHNICI DE CLASIFICARE

Modul în care instrumentele data mining analizează datele, şi tipul de informație pe care îl oferă, depind de tehnicile care sunt folosite

Clustering-ul este o operație necontrolată. Este folosită acolo unde se doreşte a se găsi grupuri de inregistrări similare în datele noastre, fără nici o altă precondiție pe care o implică acea asemănare. Această metodă statistică este folosită pentru a grupa date multi-dimensionale (adică „puncte" ce reprezintă cazuri sau observații) în grupe (*clusters*) definite algoritmic. Această metodă este utilă pentru sumarizarea unor cantități mari de informație, fiecare grupă reprezentând mai multe puncte având caracteristici similare. Clusterele distincte nu se suprapun (adică sunt disjuncte).

De fapt, analiza clasificării constă dintr-o colectie de algoritmi ce exploatează mai multe euristici fundamentate în principal pe experienta noastră „vizuală" în gruparea punctelor în „nori de puncte". Clusteringul nu necesită identificarea în prealabil a unei variabile țintă. Un algoritm verifică grupările potențiale din mulțimea datelor şi încearcă să obțină o delimitare optimă a articolelor bazându-se pe acele grupări

Clusterele sunt de obicei plasate în jurul unui centru sau valoare medie. Modul inițial de definire şi reglare a centrelor variază de la algoritm la algoritm. O metodă este de a începe cu o mulțime aleatoare de centre, care sunt apoi reglate, adăugate şi eliminate pe măsură ce analiza avansează [3].

Nearest Neighbour (cel mai apropiat vecin). Cel mai apropiat vecin reprezintă o tehnică de prezicere potrivită pentru modelele de clasificare. Spre deosebire de alți algoritmi predictivi, datele de intrare nu sunt scanate sau procesate pentru crearea unui model, ele insele reprezentând modelul. Când se prezintă un nou caz (o nouă instanță a modelului), algoritmul caută în toate datele pentru a găsi o submulțime de cazuri care se aseamană cel mai mult el, şi il foloseşte pentru a prezice consecința.

Cum metoda se bazează pe conceptul de distanță, şi acestă necesită o metrică pentru a determina distanțele. Toate metricile trebuie sa aibă ca rezultat un numar specific în ceea ce priveşte comparațiile. Oricare metrică este folosită este atât arbitrară cât şi extrem de importantă. Este arbitrară pentru că nu există nici o definiție de control a ceea ce reprezintă o metrică bună. Este importantă deoarece alegerea unei metrici afectează mult prezicerile. Aceasta inseamnă că este nevoie de un expert în domeniu pentru a ajuta la determinarea unei metrici bune.

Arborele de test e dezechilibru. Extrage în principiu sub forma unui subşir şi a unor arbori de prefix informații privind recombinarile din cadrul populației. Informația este utilizată pentru a localiza fragmentele moştenite potențial de la un fondator de boală obişnuită şi să mapeze genele bolii în fragmentul cel mai asemanator.

Data mining în genetică. Haplotype Pattern Mining – este prima abordare prin tehnologie data mining utilizată în localizarea genelor. Este foarte puternică în localizarea genelor de susceptibilitate în cazul afectiunilor umane multifactoriale. Metoda utilizează un algoritm eficient de data mining pentru a căuta pattern-uri frecvente asociate trăsăturii de interes. Spre deosebire de multe alte metode, HPM nu necesită omului de ştiintă (cercetătorului) să specifice în mod explicit modelul bolii. HPM a fost extins la fenotipuri cantitative ajustând (aplatizând) covariantele şi utilizând date pure de genotip în locul datelor cantitative de fenotip În plus, metoda prezintă robustete ridicată la date lipsă sau date eronate [4].

CONCLUZII

Datorită creşterii influentei Tehnologiei Informației şi Comunicațiilor în lumea modernă, recent au fost imaginate metode noi în Data Mining.

Atât Data Mining, cât şi Bioinformatica îşi lărgesc granitele cercetărilor atât prin dezvoltarea metodelor de laborator pentru culegerea şi analiza a biodatelor, cât şi prin dezvoltarea algoritmilor, tehnicilor şi metodelor de Data Mining

BIBLIOGRAFIE

[1] **W. Frawley, G. Piatetsky-Shapiro, C. Matheus**. *Knowledge discovery în databases: An overview*. All Magazine, vol. 13 n. 3 (57 - 70). 1992.

[2] http://www.spiruharet.ro/sesiuni-comunicari/word/5.4.pdf

[3] http://www.univermed-cdgm.ro/dwl/DoctCursS_2007.pdf

[4] **Laetitia Jourdan, Clarisse Dhaenens, El-Ghazali Talbi,** *A Genetic Algorithm for Feature Selection în Data-Mining for Genetics*, MIC'2001 - 4th Metaheuristics International Conference, p. 29-33.