Applications domain driven data mining methodology in bioinformatics

Yadan Li\textsuperscript{1}, Qinghua Bai\textsuperscript{1}, Zhicheng Chen\textsuperscript{2}
\textsuperscript{1}Tongji University, Siping Road 1239, Shanghai, (CHINA)
\textsuperscript{2}Shanghai Finance University, Shanghai, (CHINA)
E-mail : liyadan2@gmail.com; tjqbai@vip.sina.com; woods@126.com

ABSTRACT

Bioinformatics is the science of managing, mining, and interpreting information from biological sequences and structures. Although data mining in the domain of bioinformatics is popular, the two areas have largely been developing separately. There is a strong and challenging need to mine for more informative and actionable knowledge in bioinformatics. To respond this requirement, this study tries to probe domain driven data mining methodology in the field of bioinformatics, and tries to provide a new thought for bioinformatics in future research.

KEYWORDS

Bioinformatics; Data mining; Domain driven data mining methodology; Actionable knowledge.
INTRODUCTION

In recent years, rapid developments in genomics and proteomics have generated a large amount of biological data. Drawing meaningful results from these data requires sophisticated computational analyses. Bioinformatics, or computational biology, is the interdisciplinary science of interpreting biological data using information technology and computer science, and etc. The importance of this new field of inquiry will grow as we continue to generate and integrate large quantities of genomic, proteomic, and other data.

A particular active area of research in bioinformatics is the application and development of data mining techniques to solve biological problems by now. Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. Examples of this type of analysis include protein structure prediction, gene classification, cancer classification based on microarray data, clustering of gene expression data, statistical modeling of protein-protein interaction, etc. However, extant Data mining is presumed as an automated process that produces automatic algorithms and tools without human involvement and the capability to adopt to external environment constraints.

We know how to classify biological sequences (SVM, Neural Nets, Decision Trees, Rules), know how to cluster biological entities (Bi-clustering, K-means, hierarchical), know how to select features (PCA, LDA, SVM-RFE), as a result, although many patterns are mined from the data set through these technologies, few are satisfied the real needs and applications of bioinformatics. For instance, many knowledge discovered by data mining technology from one set could not help another sample.

Bioinformatics is the science of managing, mining, and interpreting information from biological sequences and structures. Advances such as genome-sequencing initiatives, micro arrays, proteomics, and functional and structural genomics have pushed the frontiers of human knowledge. In a sense, human knowledge cannot involved in the process of data mining process usually.

In addition, data mining has been advancing in strides in recent years, with high-impact applications from marketing to science. Although researchers have spent much effort on data mining for bioinformatics, the two areas have largely been developing separately.

Accordingly, in order to generate actionable knowledge satisfied genuine needs of bioinformatics completely, a great potential to increase the interaction between data mining and bioinformatics, this study tries to probe domain driven data mining methodology in the field of bioinformatics, and tries to provide a new thought for bioinformatics in future research.

RELATED WORK

Bioinformatics

The term bioinformatics was coined by Paulien Hogeweg in 1979 for the study of informatic processes in biotic systems. It was primarily used since late 1980s has been in genomics and genetics, particularly in those areas of genomics involving large-scale DNA sequencing.

Bioinformatics can be defined as the application of computer technology to the management of biological information. Bioinformatics is the study of applying computational methods to large amount of biological information in order to facilitate in biology and medicine. It has been mainly fueled by advances in DNA sequencing and mapping techniques. Over the past few decades rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. The primary goal of bioinformatics is to increase the understanding of biological processes. Some of the grand area of research in bioinformatics includes:

Sequence analysis
Sequence analysis is the most primitive operation in computational biology. This operation consists of finding which part of the biological sequences are alike and which part differs during medical analysis and genome mapping processes.

**Genome annotation**

In the context of genomics, annotation is the process of marking the genes and other biological features in a DNA sequence.

**Analysis of gene expression**

The expression of many genes can be determined by measuring mRNA levels with various techniques such as microarrays, expressed cDNA sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signature sequencing (MPSS), or various applications of multiplexed in-situ hybridization etc.

**Analysis of protein expression**

Gene expression is measured in many ways including mRNA and protein expression, however protein expression is one of the best clues of actual gene activity since proteins are usually final catalysts of cell activity. Protein microarrays and high throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample.

**Protein structure prediction**

The amino acid sequence of a protein (so-called, primary structure) can be easily determined from the sequence on the gene that codes for it. Knowledge of this structure is vital in understanding the function of the protein. For lack of better terms, structural information is usually classified as secondary, tertiary and quaternary structure. Protein structure prediction is one of the most important for drug design and the design of novel enzymes.

**Comparative genomics**

Comparative genomics is the study of the relationship of genome structure and function across different biological species. Gene finding is an important application of comparative genomics, as is discovery of new, non-coding functional elements of the genome.

**Modeling biological systems**

Modeling biological systems is a significant task of systems biology and mathematical biology. Computational systems biology aims to develop and use efficient algorithms, data structures, visualization and communication tools for the integration of large quantities of biological data with the goal of computer modeling.

**High-throughput image analysis**

Computational technologies are used to accelerate or fully automate the processing, quantification and analysis of large amounts of high-information-content biomedical images. Modern image analysis systems augment an observer’s ability to make measurements from a large or complex set of images. Biomedical imaging is becoming more important for both diagnostics and research.

**Data mining**

Data mining is often referred to as the process of extracting interesting information or patterns from data. Data mining is not specific to any industry. It requires intelligent technologies and the willingness to explore the possibility of hidden knowledge that resides in the data.
Data Mining approaches seem ideally suited for Bioinformatics (The process of knowledge discovery can be seen in Figure 1), since it is data-rich, but lacks a comprehensive theory of life at the molecular level. There are many methods of data mining shown as follows:

**Classification**
Classification is learning a function that maps a data item into one of several predefined classes.

**Estimation**
Given some input data, coming up with a value for some unknown continuous variable.

**Prediction**
Same as classification except that the records are classified according to some future behavior.

**Association rules**
Discover the high frequency pattern and discover which things appear frequently and simultaneously.

**Clustering**
Segmenting a population into a number of subgroups or clusters.

**Visualization**
Representing the data using visualization techniques.

While there are countless researchers, especially recent researchers, working on designing efficient data mining technique and algorithms. Data mining is a data driven trial-and-error process, aids to extract patterns in data without human involvement. Knowledge discovery overemphasized by innovative algorithm-driven research can not meet the needs of real world. For the domain of Bioinformatics, human knowledge involved in the process of mining is very important. For instance, learn human understandable rules that can define the epigenetic process in cancer and embryonic stem cell.

![Figure 1: The process of knowledge discovery for bioinformatics](image-url)
Nowadays, the primary challenges are moving data-driven to domain-driven and focus on discovering interesting and actionable knowledge. A new methodology on top of the traditional data-centered pattern mining framework, is called Domain Driven Data Mining Methodology.

**DOMAIN DRIVEN DATA MINING METHODOLOGY**

Domain Driven Data Mining Methodology targets to overcome three types of contradiction existing in traditional data mining. The first one is the rule vs. interestingness, that is, discovered rules are not useful or interesting to user, as well as the rules are not desired; the second is the rule vs. actionability, that is, there is a gap between rule and real world applications; the last is the rule vs. data, it can be discussed from the following two aspects, one is that data don’t contain the information that user required usually, the other is many useless or bad data, which cover the values of meaningful data. To deal with the contradiction mentioned, Domain Driven Data Mining Methodology caters for the effective involvement of intelligence, such as domain knowledge and expert experience, surrounding actionable knowledge discovery in meeting real world needs.

**Attributes of domain driven data mining methodology**

Generally, data-driven data mining system can generate a glut of knowledge, most of which are of no interest to the experts. Domain Driven Data Mining Methodology is moving data-driven to domain driven and focuses on discovering interesting and actionable knowledge. Thus, some attributes, such as interestingness, actionability and domain knowledge [7,10], need to be involved during the process of knowledge discovery.

**Attribute 1 Interestingness**

The knowledge discovered is unexpected or desired to the decision makers. It is strongly dependent on the application domain, expert knowledge, as well as experience.

**Attribute 2 Actionability**

It refers to the knowledge mined can suggest concrete and profitable action to the decision makers. It is actually relied on the application domain.

**Attribute 3 Domain knowledge**

Domain knowledge, the knowledge which is valid and used directly for a pre-selected domain of human behavior and experience or an autonomous computer activity. It is dependent on domain expert, user, environment, context, etc.

**Combined rule mining**

Combined rule mining is employed to find more actionable knowledge usually. A combined rule is composed of multiple heterogeneous itemsets from different datasets. Combined patterns take the forms of combined association rules, combined rule pairs and combined rule clusters [5], which are defined as follows.

**Definition 1 Combined association rule**

Assume that there are \( m \) database \( D_i \) \((i=1,\ldots,m)\). Assume \( I_i \) to be the set of all items in database \( D_i \) and \( \forall \ a \in I_i \neq j, I_i \cap I_j = \emptyset \). A combined association rule \( R \) is in the form of

\[
A_1 \land A_2 \land \cdots \land A_K \rightarrow T
\]  

(1)

Where \( A_i \subseteq I_i \) \((i=1,\ldots,m)\) is an itemset in dataset \( D_i \), \( T \neq \phi \) is a target item or class and \( \exists i \neq j, A_i \neq \phi \), \( A_j \neq \phi \).
Definition 2 Combined Rule Pair

Assume that \( R_1 \) and \( R_2 \) are two combined rules and that their left sides can be split into two parts, \( U \) and \( V \), where \( U \) and \( V \) are respectively itemsets from \( IU \) and \( IV \) (\( I = \{i\} \), \( IU \subset I \), \( IV \subset I \), \( IU \neq \phi \), \( IV \neq \phi \) and \( IU \cap IV \neq \phi \)). If \( R_1 \) and \( R_2 \) share the same \( U \) but have different \( V \) and different right sides, then they build a combined rule pair \( P \) as

\[
P := \begin{cases} 
R_1 : U \land V_1 \rightarrow T_1 \\
R_2 : U \land V_2 \rightarrow T_2 
\end{cases}
\]

(2)

Where \( U \neq \phi \), \( V_1 \neq \phi \), \( V_2 \neq \phi \), \( T_1 \neq \phi \), \( T \neq \phi \), \( U \cap V_1 \neq \phi \), \( U \cap V_2 \neq \phi \), \( V_1 \cap V_2 \neq \phi \) and \( T_1 \cap T_2 \neq \phi \).

Definition 3 Combined Rule Cluster

A combined rule cluster \( C \) is a set of combined association rules based on a combined rule pair \( P \), where the rules in \( C \) share a same \( U \) but have different \( V \) in the left side.

\[
C := \begin{cases} 
U \land V_1 \rightarrow T_1 \\
U \land V_2 \rightarrow T_2 \\
\vdots \\
U \land V_n \rightarrow T_n 
\end{cases}
\]

(3)

Where \( U \neq \phi \); \( \forall i \), \( V_i \neq \phi \), \( U \cap V_i \neq \phi \); and \( \forall i \neq j \), \( V_i \cap V_j = \phi \).

Based on traditional Support, Confidence and Lift, two new lifts are designed as follows for measuring the interestingness of combined association rules\(^8\).

\[
\text{Lift}_U(U \cap V \rightarrow T) = \frac{\text{Conf}(U \cap V \rightarrow T)}{\text{Conf}(V \rightarrow T)} = \frac{\text{Lift}(U \cap V \rightarrow T)}{\text{Lift}(V \rightarrow T)}
\]

(4)

### TABLE 1 : Dataset

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>stem cells</th>
<th>Disease</th>
<th>Patient ID</th>
<th>stem cells</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P_3, P_4</td>
<td>Y</td>
<td>1</td>
<td>P_3</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>P_3, P_5</td>
<td>N</td>
<td>2</td>
<td>P_4, P_5</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>P_3, P_4, P_5</td>
<td>N</td>
<td>3</td>
<td>P_3, P_4, P_7</td>
<td>Y</td>
</tr>
<tr>
<td>3</td>
<td>P_5, P_4</td>
<td>N</td>
<td>4</td>
<td>P_2, P_4</td>
<td>N</td>
</tr>
</tbody>
</table>

### TABLE 2 : Patient demographic data

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Gender</th>
<th>Patient ID</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F</td>
<td>2</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>4</td>
<td>M</td>
</tr>
</tbody>
</table>

\[
\text{Lift}_V(U \cap V \rightarrow T) = \frac{\text{Conf}(U \cap V \rightarrow T)}{\text{Conf}(U \rightarrow T)} = \frac{\text{Lift}(U \cap V \rightarrow T)}{\text{Lift}(U \rightarrow T)}
\]

(5)

Based on the above two new lifts, the interestingness of combined association rules is defined as
\[ I_{\text{rule}}(U \cap V \rightarrow T) = \frac{\text{Lift}_U(U \cap V \rightarrow T)}{\text{Lift}(U \rightarrow T)} \] (6)

\( I_{\text{rule}} \) indicates whether the contribution of \( U \) (or \( V \)) to the occurrence of \( T \) increases with \( V \) (or \( U \)) as a precondition. The value of \( I_{\text{rule}} \) falls between 0 and infinity. When \( I_{\text{rule}} > 1 \), the higher \( I_{\text{rule}} \) is, the more interesting the rule is.

For a rule cluster \( C \) composed of \( n \) combined association rules \( R_1, R_2, \ldots, R_n \), its interestingness is defined as

\[ I_{\text{cluster}}(C) = \max_{i \neq j, R_i, R_j \in C, T_1 \neq T_2} I_{\text{pair}}(R_i, R_j) \] (7)

**A NUMERICAL EXAMPLE**

A simplified numerical example is used to demonstrate the implementation procedure of domain-driven data mining. Take identification of disease between parts and gender as an example. There are eight parts of stem cells (\( P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8 \)). Suppose human understandable pattern is a key factor in the identification of patient. This example only considers patient demographic dataset and the number attributes of dataset (See TABLES 1 and 2). The concrete procedure can be summarized as follows:

- Assume \( \text{MinSupport}_1 = 0.2 \) and \( \text{MinConfidence}_1 = 0.4 \). Thus, single association rules can be shown as follows: \( F \rightarrow Y, F \rightarrow N, M \rightarrow Y, M \rightarrow N, P_3 \rightarrow Y, P_3 \rightarrow N, P_4 \rightarrow Y, P_4 \rightarrow N, P_5 \rightarrow Y \) and \( P_5 \rightarrow N \).

**TABLE 3 : Combined association rules**

<table>
<thead>
<tr>
<th>Rules</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift_1</th>
<th>Lift_2</th>
<th>Lift_3</th>
<th>( I_{\text{rule}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F \land P_4 \rightarrow Y )</td>
<td>0.3</td>
<td>0.75</td>
<td>1.5</td>
<td>1.3</td>
<td>1.25</td>
<td>1.0</td>
</tr>
<tr>
<td>( F \land P_5 \rightarrow N )</td>
<td>0.2</td>
<td>0.67</td>
<td>1.3</td>
<td>1.1</td>
<td>1.7</td>
<td>1.4</td>
</tr>
<tr>
<td>( M \land P_3 \rightarrow Y )</td>
<td>0.2</td>
<td>0.67</td>
<td>1.3</td>
<td>1</td>
<td>1.7</td>
<td>1.3</td>
</tr>
<tr>
<td>( M \land P_4 \rightarrow N )</td>
<td>0.2</td>
<td>0.67</td>
<td>1.3</td>
<td>1.3</td>
<td>1.1</td>
<td>1.1</td>
</tr>
</tbody>
</table>

**TABLE 4 : Combined rule pairs**

<table>
<thead>
<tr>
<th>Pairs</th>
<th>Combined rules</th>
<th>( I_{\text{pair}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair_1</td>
<td>( F \land P_3 \rightarrow Y )</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>( F \land P_4 \rightarrow N )</td>
<td>1.3</td>
</tr>
<tr>
<td>Pair_2</td>
<td>( M \land P_4 \rightarrow Y )</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>( M \land P_4 \rightarrow N )</td>
<td>1.1</td>
</tr>
<tr>
<td>Pair_3</td>
<td>( F \land P_4 \rightarrow Y )</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>( F \land P_5 \rightarrow N )</td>
<td>1.4</td>
</tr>
</tbody>
</table>

- Assume \( \text{MinSupport}_2 = 0.2, \text{MinConfidence}_2 = 0.6 \). According to equations (1, 4, 5, 6), combined association rules can be shown in TABLE 3.
- Based on equations (2, 3, 7), combined rule pairs can be shown in TABLE 4.
- Finally, The actionable patterns indicate that \( P_3 \) is the key part leading to the disease of male patient. Meanwhile, \( P_4 \) is the key part leading to the disease of female patient.

**CONCLUSION**
Bioinformatics and data mining are developing as interdisciplinary science. Data mining approaches seem ideally suited for bioinformatics, since bioinformatics is data-rich but lacks a comprehensive theory of life’s organization at the molecular level.

However, although many patterns are mined from the data set through data mining technologies, few are satisfied the real needs and applications of bioinformatics. For instance, the existing problem is that the domains of expertise present amongst potential users, so it can be difficult for the database curators to provide access mechanisms appropriate to all. Since data mining applications have urgent requirements for discovering actionable knowledge to decision makers in bioinformatics, datacentered traditional data mining cannot satisfy the needs fully. Thus, DDDM is developed to provide a systematic overview in acquiring actionable knowledge through human mining system involved ubiquitous intelligence such as expert intelligence and human intelligence.

To conclude, this study proposes DDDM towards the fields of bioinformatics for actionable knowledge. In particular, a numerical example is employed to verify its effectiveness. Furthermore, more experiments will be confirmed in future research.

REFERENCES