A scalable data mining architecture for bioinformation

R. Li¹, Z. Zhang¹, S. Cao¹, Y. Zhu¹,² & Y. Li²
¹Department of Computing and Information Technology, Fudan University, China
²Shanghai Centre for Bioinformation Technology, China

Abstract

Bioinformatics is a new field of data mining research, and data mining is also a promising tool for bioinformatics. Research has been done on how to bridge these two attractive fields. With the critical progression on genetics and high-throughput biotechnologies, all kinds of bio-data have been explosively produced and accumulated. With the success of global views of DNA sequences, gene expression levels, etc. on the genomic scale, now it is possible to discover the nature of life and to promote biology and medicine research further. However, to achieve these goals, accurate large-scale data processing technologies are required. Data mining has been known as a powerful tool for this purpose, but it is still difficult to develop and apply a data mining system for different analysis functions. In this paper, a new scalable data mining architecture for bioinformation analysis, "The Architecture of Bioinformation Data Mining Application Platform", has been proposed to facilitate biologists use of complex data mining technology and to develop a professional data mining system by conveniently developing, customizing or trimming a data mining system for some specific bioresearch. Logically, the architecture is composed of three tiers: a data mining algorithm tier, a analysis logic tier and a profession application tier, with the mapping of data mining algorithm-analysis model-special application as backbone. The architecture can be implemented by using the distributed client/server model. A gene expression data mining system BioMiner is designed and implemented to perform and validate the efficiency of the architecture.
Introduction

1.1 Bio-data analysis

In recent years, revolutionary progression in bioscience makes it more and more important to process, analyse and mine bio-data to shape bioinformation knowledge. Bio-data has been produced and accumulated explosively, it provides the data basis to discover gene regulation network, life evolution and other important bio-problems, but only a small part of bio-data has been processed. Bio-data is characterized as massive, heterogeneous with high dimensions and net character in nature; its analysis is beyond the ability of traditional analysis methods and decision supporting technologies. Data mining begins to play a more important role because of its excellent ability in the large-scale data analysis.

After more than ten years’ development, research on data mining algorithm has been matured, and its research focuses have been transferred from algorithm to application combined with characters of application areas (Piatetsky-Shapiro, [1]). Data mining has gained more attention and deeper application on various industries. Combined with the feathers of each industry, data mining flourishes in many industries, such as CRM (Customer Relation Management), electronic business and finance etc.

In the area of bioinformation process, research method of modern biology has changed profoundly with the broadening information technology, and it more depends on the clues and bases deduced from the analysis results of information technology. Powerful data processing tool becomes the key for modern bio-research development, and the support from large-scale data analysis technology such as data mining is urgently requested.

There are only several kinds of data mining methods (Han and Kamber, [2]), such as clustering, classification, association rule and outlier detection, etc., but they can achieve various analysis functions. Each method has many algorithms and each algorithm has its own characteristics. Usually the same algorithms can be used to solve different bio-problems. It is very important for bioinformatics to research on how to use the limited kinds of methods to solve various bio-problems.

At the beginning, only clustering method was applied in bio-data analysis and it has become the most common tool and lots of research has been done on it. Now more data mining methods have been introduced into bio-data process, such as association rule (Creighton and Hanash, [3]) etc, and various data mining algorithms have been adapted to bio data process to develop new data mining systems. However, most of these systems are designed only for specific problems. For various bio-research analysis systems, which have specific analysis functions, a proper architecture is required to help custom a data mining system, and the research on bioinformatics software architecture becomes an urgent task.
1.2 Data mining architecture research

At the initial stage of data mining process, most software tools came only with single data mining algorithm. This kind of tools requires user master data mining technology and carry out lots of work on data preprocessing.

However, in reality, users usually need several kinds of algorithms at the same time and a lot of work need to be done on data cleaning and data preparation. Around 1995, the second generation of data mining system named Data Mining Tool Set was proposed. Tool set provides more algorithms such as association rule, classification and clustering, etc., data transformati0n, and visualization methods. This kind of softwares are called Horizontal Data Mining Tools since they are not application-oriented (Kohavi and Sahami, [4]).

Horizontal Data Mining Tools save the users’ work on data cleaning and data preparation, but the developers and users still need to master data mining technology. Their application space is also very limited.

Vertical Solutions were developed to further promote data mining application, solution providers began to produce complete solutions for specific applications, such as KD1, Options&Choice, HNC and Unica Model 1, etc.. This kind of solutions were designed for users in these areas to use given softwares to solve their specific problems. Applications of Vertical Solutions were limited due to their narrow application space and lack of flexibility.

To provide the vertical solution, a complete system must be designed for every practical problem. In reality, many process and algorithms are the same or share some common properties, we can abstract these common parts, and use them as a basis for different applications. Also we will provide a configuration tool to construct analysis flow and parameters for different problems. It is just like constructing independent vertical solutions on the common basis.

1.3 Data mining research for bioinformatics

Data mining systems used in bioinformatics usually can be divided into two kinds: general data mining software and bioinformatics-oriented software. The former is generally Horizontal Data Mining Tool, the other such as Cluster&TreeView designed for gene expression data analysis is Vertical Solution. Their system architectures cannot help bioinformatics research effectively.

In bioinformation data mining research, covered issues mainly focus on algorithm development as well as its application, but how to produce an analysis tool that is efficient and easy to use still remains as a problem. When software is requested by a particular research, the researchers have to spend lots of time trying to develop a complete data mining system, but it is complex. It is still difficult to apply data mining. It is time to research on the data mining architecture for bioinformation to promote data mining application in bioinformatics and to facilitate bio-researchers to use data mining technology readily.

In this paper, bridging researches on data mining software architecture and the feathers of bio-data analysis, a scalable multi-tier architecture for
bioinformation data mining application platform is proposed, which is inserted
with analysis logic tier between data mining algorithm tool tier and data mining
application tier and tightly combines data mining technology with analysis
application. Therefore, on the one hand, it makes the users overcome the
difficulties of using data mining algorithms and devote themselves to their
professions; on the other hand, it can be adapted to meet different requirements
of different industries with analysis models in analysis logic tier.

Data mining technology has achieved a remarkable success in many areas
such as CRM, etc. in that it summed up typical issues in these areas and
discovered flows to solve these issues with data mining technology. The
proposed architecture provides the mechanism for the developer to configure and
envelop the flows for extracted interesting issues.

On the part of the System and methods, we introduce the concepts in the
architecture and hibercarchy of the architecture, then as an example of using this
architecture, a scalable client-server data mining system – gene expression data
mining system BioMiner, is described, and its analysis models and major
implementation techniques are introduced. At last, our discussion is focused on
some future work.

2 System and methods

In this paper, a three-tier data mining application platform architecture is
designed for bioinformation analysis, and a new-type application platform model
is proposed for data mining system. Figure 1 shows the hibercarchy of the data
mining architecture, in which gene expression analysis is used as an example of
analysis logic tier.

2.1 Concepts in the architecture

*Data mining algorithm tool* consists of an algorithm and its pre-process,
such as data transformation required by the algorithm, and is a complete code
unit. An algorithm tool may be used for different analysis models, thus flexible
parameters are necessary. For usability issue, default parameters are provided;
user also has the option to customize these parameters.

XML can be used to describe parameters, which makes the algorithm interface
easier to understand.

*Analysis logic* is an analysis function set for a certain analysis direction,
including one or more analysis models. An analysis logic may be corresponding
to several application systems.

*Analysis model* is a single and practical analysis function that can be
fulfilled by one algorithm or a process flow that composes of more algorithms or
even several kinds of algorithms including result visualization methods, and
enveloped as a process module for a practical analysis task. An analysis function
can be achieved by different algorithms, and different algorithms can obtain
different effects that can be assessed by some figures, such as *figure of merit*
(FOM) (Yeung et al. [5]), so the results need to be compared automatically to
output the optimal result.
Analysis model is just a unit of analysis function with understandable description. It can meet requirements from different research areas and become more standard through the application.

**Application system** is a software system as a representation of analysis logic on a certain research direction, and a GUI (Graphic User Interface) where data mining analyses are implemented. An application system is corresponding to an analysis logic.

**Analysis model configuring** is a procedure to set up the mapping of algorithms to analysis models and related operations. For computing technology, a data mining algorithm is an analysis tool, and as for practical application, it can complete various tasks. The analysis model configuring procedure fulfils the mapping and envelops required algorithms and work flows.

**Application configuring** is a processing procedure to produce application system corresponding to the analysis logic according to the certain professional research direction.

**ETL (Extract, Transform and Load) tool** is a tool software for data extracting, transforming and loading.

![Figure 1: The hiberarchy of the data mining architecture for Bioinformation. The application and analysis logic tier are shown by the example of gene expression microarray data mining system.](image)

### 2.2 Hiberarchy

The architecture takes data mining algorithm tool as its core, and inserts analysis logic tier between data mining algorithm tier and application system tier.

On the analysis logic tier, logic models needed by the specific analysis logic are extracted, and mapped to proper data mining algorithms and analysis flows. Therefore, for users, data mining algorithms and their practical
applications are isolated from practical application, and users can understand the
analysis models of an analysis logic and need not understand data mining
technologies supporting below. The users’ views are only related with their own
researches. Most parameters are characters of bio-data. Users can get rid of being
restrained by professional computer knowledge and devote themselves to their
professional areas with the technological support of the software system.

3 Implementation

Research on bioinformatics was focused on the analysis of sequence data. Recently, it is transferred to the backward engineering research presented by
analysis of gene expression data from microarray, whole genome association
discovery and gene regulatory network (Califano, [6], Pennisi, [7]). CDNA
microarray is a very promising high-throughput technology. Tens of thousands
of gene expression levels can be monitored by one microarray experiment.
Changes in gene expression levels during life activities are the core problem of
bioresearch. Their analysis is the typical data mining application in
bioinformatics. So we set it as an example of implementation of the architecture
proposed in the paper, design and implement a gene expression data mining
system to practice and validate the architecture. We name it BioMiner.

3.1 Gene expression data mining system: BioMiner

In the later part, we will discuss BioMiner’s system architecture, analysis models.

3.1.1 System architecture of BioMiner

Using client/server model, BioMiner has 2 clients as in figure 2.

Database and User Management client is responsible for extracting and
transforming outer data from LIMS (Laboratory Information Management
System), other databases or flat files into BioMiner’s database by using ETL tool.
It also manages microarray database, data mining model database and users.

The analysis client provides a panel for analysis functions, where the
analysis requests of the users can be satisfied.
3.1.2 Analysis models: gene expression data analysis functions

Analysis models are extracted from practical requirements and present analysis ability of the analysis logic.

The main analysis models of BioMiner are described as followed:

3.1.2.1 Gene expression similarity analysis model

Similarity of measurable data such as sequences, gene expressions, etc., implicates the similarity of their functions. This commonality is very important for function genome research. The functions of genes whose expressions are similar are in the same types (Eisen et al., [8]).

Clustering method divides the genes into groups (clusters) according to their expression similarity. Different clustering algorithms have different sensitivities to different types of data (Yeung et al., [5]), so algorithm should be chosen by algorithms’ effect.

3.1.2.2 Gene expression similarity explanation model

Using classification techniques such as the decision tree, neuron network analyzes the results of clustering to obtain expression features of each cluster and orders conditions by their differentia levels.

3.1.2.3 Gene expression concurrence analysis model

With the association rule method, we can predict the relationships of interaction and reciprocity of genes. For example, an association rule denoted as $A \uparrow \Rightarrow B \downarrow C \uparrow$ means that, when expression level of gene $A$ increases, generally at the same time that of gene $B$ decreases and that of gene $C$ increases (Creighton and Hanash, [3]).

3.1.2.4 Special gene expression analysis model

By using outlier detection method on the gene’s expression data under different conditions, it is expected to find a certain gene expression profile under some condition, significantly different from that under other conditions. For example, it will be helpful to find the features of some diseases if we can identify some gene’s expression level only change under rare conditions.

3.1.2.5 Gene expression pathway analysis model

With time sequence pattern discovery technology, we can mine gene expression profiles on time series. These rules can be used to analyze the effect of medicines, the relationships of gene expressions on time series, etc. After genes are supplied with some medicines, by testing and analyzing the gene expression levels of the experiment objects on a series of time points, gene expression waving rule of the whole genome, gene on/off path and gene regulation series, etc., can be discovered. If some joint points on the gene expression pathway of some disease can be discovered, then it is possible for doctors to control the joint points according to the stage of patients’ symptoms (gene expression profile), thus prevent the diseases from developing to the next stage.

3.1.2.6 Gene expression bypath analysis model

By analyzing gene expression data that are special and in small size, we can gain genes’ special
expression profile in time sequence. It is very important for some conditions such as warning of medicine side effect, minor patients’ medicine reaction, etc.

It can be implemented by several ways. One is to use some density-based clustering algorithm to discover the small data set, then use time sequence pattern method to discover those expression profile; another method is using classification technology such as decision tree algorithm and its visualization method, to analyse each cluster’s profile, then the major joint points of all clusters and expression profiles can be shown.

With microarray’s application broadening in various industries, new analysis logics and analysis models will be developed by researchers continuously.

3.1.3 Implementation techniques

Under its multi-tier architecture, BioMiner employs component software technology, supporting PMML (Predictive Model Markup Language) and XML standards on data source, data mining procedure and result presentation.

Data mining normalization is an ongoing project. W3C (World Wide Web Consortium) has agreed PMML as data mining language standard. Bioinformation data is huge, and mostly stored in heterogeneous databases so it is difficult to present net character of bioinformation. Standardization is one of the first problems encountered in bioinformatics and a basis for many other problems. XML is shaping up to be the standard computer language for bioinformatics (Ezzell, [9]).

For example, the parameter definition is described as an XML string under DTDs (Document Type Definition) for each algorithm, so the meanings of the parameters are clear, and each algorithm interface is simple but more understandable.

We choose JAVA as developing language and use EJB (Enterprise Java Bean) to implement data mining algorithm and analysis model. JAVA has many merits, such as robust, safe, easy to use and cross-platform ability. EJB technique supports the multi-tier architecture in this paper well. EJB technique combines with object technique and component software technique and has been widely used in component software development. EJBs can be distributed on Local Area Network (LAN) and Internet, run cross-platform. For finish their work, users only need to understand EJBs’ interfaces and configure their parameters without the inner of these components.

By using EJB, it is easier to extend the system’s functions and force on bioresearch process. It can keep the same interface, but using different algorithm to implement the functionality. This makes easier for introducing new or better algorithms to meet requests, and keep system structure untouched. On analysis logic tier, it is convenient to construct new analysis models to extend application areas and adapt it into professional analysis systems for different applications.

BioMiner uses IBM DB2 as backbone database, JDBC is used to connect database. JAVA combined with JDBC can realize write once, use everywhere.
3.2 System features

BioMiner, based on the multi-tier architecture and component software technology, isolates data mining algorithms and their applications in the user’s view. Because of JAVA and EJB technique, it can be distributed to different operation platform, scalable and easy to use and trim, so it affords facilities for system extension and secondly development.

The followings features sum up the architecture design, analysis model setting, and implementation technique choosing.

3.2.1 Multi-tier architecture

Based on the architecture, technologically, BioMiner makes data mining technology combined with practical application tightly, but for the users, they are isolated from detailed implementation. Users can understand the functions of the system on the view of analysis logic level and do not need grasp data mining technology. On the data mining algorithm tool tier and analysis logic tier, each entity (data mining algorithm tool, analysis model and analysis logic) is generic. An algorithm tool can be used in several analysis models, an analysis model in several analysis logics and an analysis logic in several application systems.

3.2.2 Scalable distributed system structure and independent entity

For the multi-tier architecture, algorithm and analysis become concepts and entities on two levels. In implementation, by using EJB technique, every algorithm tool and analysis model is distributed as an individual unit independently. They can be removed from the system easily and new entities can be embedded into the system seamlessly. A data mining system can be extended or trimmed for practical application properly.

3.2.3 Independence to operation platform

Owing to JAVA’s cross-platform advantage, BioMiner is independent of operation platform so that it can run in Windows environment and in LINUX as well.

3.2.4 Enriched analysis functions

BioMiner is configured with various gene expression analysis functions including current expectation on typical gene expression data mining analysis. It can easily insert new analysis function.

3.2.5 Parallel, easy to use

JAVA multi-threat technique provides parallel to speed up algorithm running. How to make data mining system friendly to users and easy to use is still an important problem for data mining research. BioMiner exerts itself to this direction too. By separating data mining algorithms with their application, users can easily use the data mining applications based on their professional interests.
4 Conclusions

In this paper, we proposed a new-type scalable data mining architecture to facilitate shaping bio-knowledge. Using multi-tier architecture, implemented by Java multi-thread, EJB technique and XML, the proposed system “BioMiner” shows good performance and usability, flexibility, and facility to customize.

The future work is to extend the architecture under grid environment. Bio-data analysis is data and computation intensive, for which grid technology provides optimal support and its application research is based on the bioinformatics as an example at first. It is proper to perform data mining under grid to enhance and expand data mining application. And we plan to optimise the gene expression data mining system BioMiner and publish it as a practical tool.

Acknowledgements

This work was supported by the National High Technology Development Program of China and the Key Program of Basic Research of Science & Technology Commission of Shanghai Municipality of China. In addition, the first three authors should be regarded as joint First Authors.

References

[2] Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques (chapter 5-8), San Francisco, CA: Morgan Kaufmann, 2000