The Research and Implementation of Data Mining Component Library System

Peng Peng  Qianli Ma  Chaoxiong Li

College of Computer Science and Engineering, South China University of Technology, Guangzhou 510640, PR China

Abstract: With the wide application of business intelligence in corporate, the demand for data mining software increases daily. To improve the efficiency and quality of the reusing data mining software and reduce the period and cost of developing data mining application system, this paper proposes a new component library system of data mining. Through componentization of data mining algorithm, this system implements varied core algorithms of data mining in the form of components. In this way, the efficiency and quality of developing data mining software are improved significantly to meet various application demands.

Key words: data mining, component library, software reuse

Data mining is a new developing technology for enterprise data and information integration. It can reduce the operation cost, increase profit, and strengthen market competition of the enterprise. Generally, there are two ways to establish a data mining application tailor to an enterprise: using business intelligence solutions and products available on the market, or developing data mining algorithms all by oneself. However, both of them are impractical in cost and time. The former one costs a lot, while the latter requires developers to be familiar with both enterprise business and data mining technology.

Software reuse is a solution to avoid repeated work in the software development. It is regarded as a viable approach to solve the software crisis and promote efficiency and quality of software production. As a kennel technique to support software reuse, software component technique gets increasingly wide attention. To fully make use of reusable component, and support mass component’s production, classification, search, assembly and maintenance, component library is very important. Applying software component technique to data mining, wrapping individual business modules of data mining in the form of components, and using component technique to achieve the organization, management and retrieval of the components, could greatly increase the reusage efficacy and quality, and decrease the cost and period of data mining application development. The demand of variability of data mining tasks could be met as well. And the application of data mining technology can be broaden[1].

Through componentization of data mining algorithm, design of data mining component library and development of component adaptive and assembly technique, this paper presents a new component library system of data mining and implements varied core algorithms of data mining in the form of components, which significantly improves the efficiency and quality of developing data mining software. By assembling components to analysis module and seamlessly integrate them into industrial information system, the organic integration of data mining software and industrial information system forms industrial information decision analysis system to meet various application demands. Compare to the existing data mining software, it can provide data mining service with more humanity characteristic and higher success rate, which is more suitable for real-time decision analysis application software.

The rest of this paper is organized as follows. In the next section, we introduce the related work. The discussion of the design of component model is given in Section 2. Section 3 gives the classification and retrieval strategies of the component library. Finally, Section 4 lists some conclusions and discusses some areas of future research.

1 Related work

Current research in software engineering has made profound study on reusable software component library system, and obtains many fruit, including some actual systems, such as REBOOT (Reuse Based on Object-Oriented Techniques), standards,
such as RIG (Reuse library Interoperability Group) and NATO (North Atlantic Treaty Organization), and Models, such as ALOAF (Asset Library Open Architecture Framework) [2], RIG is a volunteer, consensus-based organization composed of members from government, academia, and industry. NATO prefers a faceted classification scheme. Each facet need be neither independent of every other facet nor applicable to every possible component [3,4]. The STARS (Software Technology for Adaptable Reliable systems) project, which is sponsored by DARPA, addresses the problems of resource sharing and seamless interoperation among component libraries, and proposes Asset library Open Architecture Framework (ALOAF) in 1992. Beside, research on component library represented by software reuse and its associated techniques in Jade Bird Project of Peking University has obtained some achievements in China. In Ninth Five-Year Plan, Jade Bird Project developed an application integrating environment called JB3, which is based on heterogeneous platform, and has multi-information sources interface. Based on the corresponding component model, it created the Jade Bird component library data model. Combining with other CASE tools, it support component description (using Jade Bird description language JB-CDL), classification, management, store, retrieval, and composite, which feed the need of software developing process based on “component-architecture” reuse.

In data mining research field, Some main data mining software, such as IBM Intelligent Miner, SAS Enterprise Miner, SPSS Clementine, etc., can provide general mining process and mining model. All of these data mining tools can finish different data mining task, but they are all indecomposable tasks. We cannot separate one or some modules and embedded into other systems. If any other enterprises want to develop their own data mining application system, they can only do it based on these tools. It consumes expensive charge and lacks of flexibility for their system.

Data mining component library based on software re-usage can resolve this dilemma. Large number of components in the component library can rapidly construct application system of different fields, which greatly increase the re-usage efficacy and system development, and reduce the cost. Few researches have deal with the data mining component library. In [5], an idea of component data mining system is presented, but the concrete research is not involved. In this paper, we implements varied core algorithms of data mining in the form of components, and use the component library to support the classification, storage, retrieval and extraction of the components. It can provide an effective component re-usage environment for data mining, which significantly improves the efficiency and quality of data mining.

2 Data mining component model

First, this section discusses the definition of data mining model and on the basis of this, designs and implements the data mining algorithm model and component model. Component model is the core of component-based development method, the essential feature of component, and the abstract description of the relation between components. Component is the basic element of component library. Each component represents a process of data mining task and implements a certain data mining function.

2.1 Definition of data mining model

This section gives the mathematics definition of data mining problem [6], which is the foundation of the definition of data mining component model.

The process of data mining usually consists of two stages:

- Training stage: build mining model based on training data.
- Application stage: apply the mining model to new data.

In train stage, consider the given training set:

$$T = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}_{f_{i+1}}^A\}, \quad (1)$$

where \(x_i\) denote point \(i\) of the \(M\) points in the \(d\)-dimension attributes space. Suppose training set \(T\) is the sample data of the unknown function, where \(f\) belongs to function space \(V\), which is defined in \(\mathbb{R}^d\), training objective is to recover function \(f\), given data set \(T\), let

$$y_i = f(x_i) \quad (2)$$

In order to get such function \(f\), we need to solve the following minimization problem:
\[
\min_{f \in V} R(f)
\]  
(3)

where \( R \) is the operation, which mapping \( V \) to \( \mathbb{R} \). In training stage, all of the data mining algorithms build mining models using training data.

### 2.2 Data mining algorithm model

Because data mining algorithm model is in the core position of the data mining process, before discussing data mining component model, we consider the design of the algorithm model first.

#### 2.2.1 Process of algorithm model

From the definition of data mining model in Section 2.1, we can abstract the algorithm model, which consists of 6 parts:

- **Input**: training data or application data for algorithm or model, which can be from different types of files or databases.
- **Setting**: set meta information and parameters of the data, which will be operated in data mining algorithm.
- **Algorithm**: extract training data by mining input, set parameters of algorithm and attributes of data by mining setting, process mining events by feedback, and last build mining model.
- **Feedback**: event processing mechanism in data mining algorithm.
- **Model**: model is built by data mining algorithm and can be applied in application.
- **Result**: mining result.

#### 2.2.2 Detail design of algorithm model

The detail design of algorithm model is shown as Fig 1:

![Fig 1 Flow of algorithm model](image)

### 2.3 Design of component model

Component is the basic element of the data mining component library. Each component stands for a process of data mining task, and implements a certain data mining function. The core design of data mining component is to wrap individual function modules of data mining business process in the form of components, and provides standard service. Based on these services, we can complete complicated data mining task by assembling components.

#### 2.3.1 Execution process of component

Although component is relative independent unit, which can provide different service, according to the abstraction of data mining process, we can see, each data mining component gets input message from external environment or previous conjoint component, processes corresponding message based on the parameter set by user, and produces intermediate result or final mining result for next component.

As we can see from Fig 2, the execution process of component includes transversal part and longitudinal part. Transversal part indicates the communication between different components. Components use intermediate results to transfer operation information, including data or model, and complete complicated data mining task by assembling components. Longitudinal part is the execution process of single component, including getting operation information from parent component, operating corresponding message based on the parameter setting, producing intermediate result, and visualizing the result [7].

![Fig 2 Execution process of component](image)

#### 2.3.2 Definition of component interface

Component must have an abstract description of the service it provides, which acts as the contract between service customer and provider. That is the interface of component.

Different components can collaborate with each other through component interface. As is shown in...
Fig 3, the left part shows different component users, and the components developed by different developers are shown in the right part. Component interface connects the users and the developers, labeled as “Interface” in the figure. The component developers only think about the function implementation of component interface, while the component users only concern the component interface which they dependence on.

![Component Interface Diagram]

**Fig 3 Component interface**

2.3.3 Design of component class

We define component class as `Component`. The function of `Component` includes getting input message from parent component, execution, giving output message to child component, and producing temporary file.

2.3.4 Design of property class

We define property class as `ComponentProperty`. It is the object used to set component parameters, display and update component parameters, and input meta information.

2.3.5 Design of result class

We define result class as `ComponentResult`. It is used to visualize the component operating result, including data structure, data recording, and model. By inheritance, we can design different types of result classes. Some are public to all the components, such as `DataComponentResult`, which is used to show structure and result of the temporary data, and some others are special classes, like `DecisionTreeComponentResult`, which is special for display decision tree.

2.4 Implementation of component model

Based on the component model in Section 2.3, we can construct the data mining component entity. Each component forms from three relevant classes, i.e. component class, property class and result class.

![Inheritance Graph]

**Fig 4 Inheritance graph of component class**

Component class is the main class of component entity. It implements the operating function of component. As we can see from Fig 4, there are 6 basic types of subclasses derived from component class, including `InputComponent`, `PrepareComponent`, `AlgorithmComponent`, `AssessmentComponent`, `DeployComponent` and `VisualComponent`. All the corresponding component classes of component entities are derived from `Component` or its 6 subclasses.

3 Classification and retrieval strategies of component library

Component library is the important establishment for software reuse. If there have been lots of different components, it is hard to retrieval appropriate component when develop a new system. So classification and retrieval technique of component is required to research. The classification and retrieval technique of component divides components into different component classes for searching and matching using component classification information. Good classification mechanism can significantly improve retrieval efficiency [8].

Classification and retrieval of component can be divided into three steps:

- User submits demand information to retrieval system.
- Translate the user demand into High-level abstract basic retrieval statement.
- Change basic retrieval statement into the retrieval statement, which can be actually executed by retrieval system.

3.1 Component classification based on data mining function

According to the abstraction and summary of the data mining business, the entire data mining business can be divided into several components with larger
Component library system provides various retrieval methods, which brings convenience for component reuse and improves retrieval efficiency. There are several retrieval methods, including property retrieval, keyword retrieval, facet retrieval, hypertext links retrieval and fulltext retrieval\[8\].

- **Property retrieval**
  Property retrieval is a simple retrieval method, which queries using the properties of component, such as name, developer, development date or input date.

- **Keyword retrieval**
  Keyword retrieval is a retrieval method that queries component, according to a certain term in the term space. If it cannot find appropriate component based on the directly related term, it can be also referred to the components belong to other related terms.

- **Facet retrieval**
  Facet retrieval can be used to search among categories. The relationship between different categories can help to find out appropriate component.

- **Hypertext links retrieval**
  Actually, hypertext links retrieval also browses among categories, as the same as facet retrieval. But different from facet retrieval, it adopts hypertext, which is easy to operate in web retrieval.

- **Fulltext retrieval**
  In component library, fulltext retrieval is suitable for retrieval of document information and source code. Using this retrieval method, it is necessary to establish a fulltext index for all the data first in the background. System will select such components related with the documents or source codes that include the words submitted by the user.

## 4 Conclusions and future research

This paper proposes the application research of component technique in data mining, and presents a new component library system of data mining. We implement varied core algorithms of data mining in the form of components, and use the component library to achieve the organization, management and retrieval of the components. Through componentization of data mining algorithm, wrapping individual business process of data mining in the form of components, the efficiency and quality of developing data mining software are improved significantly to meet various application demands, and thus the application of data mining technology can be broaden.

Future work include design a knowledge library, which uses knowledge to describe the data mining component, to support the data mining component library, a rule library, which can modify flexibly and easily, a scheme library based on CRISP-DM data mining process, and a key-word library for
component retrieval. By using the inference engine of the knowledge library and the rule library, we can design the recommendation strategies for data mining components. It realizes the control of knowledge library in the application and development process of data mining.

Acknowledgements

This work is supported by Natural Science Foundation of Guangdong Province, China (07006474) and Science and Technology Planning Project of Guangdong Province, China (2007B010200044). (2008国家大学生创新性实验计划项目资助)

References

[7] AlphaMiner node design criterion. HIT- HKU Business Intelligence Joint Laboratory, 2006