Improved Fuzzy Hierarchical Clustering Algorithm and its Application in Software Re-architecting

Sheng Gao, Chenghua Lu and Ning Miao

Department of Information Science and Technology, Pearl River College of Tianjin University of Finance and Economics, Faculty of Data Engineering, Xiangru street, Jingjin new town, Baodi District, Tianjin, China
Email: gslypswu@163.com

Abstract. Computer technology changes with each passing day. As a member of software system, the architecture of legacy system becomes complex and difficult to adapt to complex software functional requirements in the process of evolution. It has not only affected the late legacy system maintenance work, but also to the enterprise has brought serious losses. In order to solve the above problems, this paper presents a Fuzzy Hierarchical Clustering algorithm Based on Information Loss (FHCBIL). There are two improvements in FHCBIL algorithm: On the one hand, this algorithm extends the weights allocation of entity features, which solves the problem that the two element relation cannot reflect the influence degree of entity feature on entity. On the other hand, FHCBIL algorithm uses the information loss as a similarity calculation method, avoiding the problem of sensitive data set to the irregular data set in the clustering process. In this paper, we further apply FHCBIL algorithm in software re-architecting, give the model of software re-architecting, and implement new nested software architecture. In order to evaluate the quality of the new architecture, we evaluate the new architecture from two aspects: the rationality of the new architecture hierarchy and the accuracy of entity division. The experimental results show that the new architecture based on FHCBIL algorithm has high cohesion and low coupling. And the accuracy of the new architecture is high.

1. Introduction

In the middle of the 1960s, with the emergence of large capacity, high speed computer and advanced language, software development increased rapidly. At this time, the scale of software system is bigger and more complex, the clarity and readability of software architecture become an important standard for software developers to solve later maintenance problems. With the rise of data mining technology at the end of the 1980s, clustering algorithms in data mining have become an important technology in software re-architecting. In 2007, Maqbool[1] and others put forward the WCA (Weight Combined Algorithm and WCA) algorithm applied in software re-architecting. But the algorithm abandoned small competitive entities in the iterative process, which result in low quality of clustering. In 2014, Masoud[2] built software re-architecting model based on hierarchical clustering, and selected classes as the basic unit of software architecture. In 2016, Mkaouer[3] allocated the weight of entities, and used hierarchical clustering method to reconstruct object-oriented software architecture, and realized the software re-architecting on the granularity level. In 2020, COMSATS University puts forward advances in software reconstruction, which is to apply neural network to software reconstruction[4]. In this study, class is regarded as the minimum unit of software reconstruction, and software re-architecting is completed with high quality. Based on the above research background, Fuzzy hierarchical clustering is applied in software architecture reconstruction, it can not only make the
reconstructed new architecture level clear, but also improve the quality of new architecture. This is of great significance for the later maintenance of the software system, especially the legacy system.

2. The Model of Software Re-architecting Based on Fuzzy Hierarchical Clustering

There are two requirements for software re-architecting: firstly, the new architecture after the software re-architecting has a high cohesive, low coupling structure. Secondly, the partition of entities in the new architecture is high accuracy. Based on the above two requirements, we build a software re-architecting system model based on fuzzy hierarchical clustering. This system is divided into three modules, which are the data extraction module of the software system, the fuzzy hierarchical clustering module and the new architecture evaluation module. Software re-architecting system model framework based on fuzzy hierarchical clustering, such as figure 1.

2.1 Data Extraction Module of Software System

The function of this module is to convert the source code of the software system and to extract the data, and to complete the preliminary data preparation in the process of the software re-architecting. The data extraction module of the software system includes the following three parts:

- Selection of legacy systems.
- By reverse engineering technology, conversion of the source code of the legacy system.
- Data conversion of UML model diagram.
2.2 Fuzzy Hierarchical Clustering Module

2.2.1 Improved strategy of FHCBIL algorithm

There are two shortcomings in the traditional hierarchical clustering algorithm: on the one hand, the relationship between the entity characteristics and the entity belong to the two binary relation. On the other hand, distance based similarity computing is not suitable for clustering among components in software re-architecting. In view of the above shortcomings, the Improved strategy of FHCBIL algorithm has two aspects:

1. Extended the weights of entity characteristics
   Improved strategy applies fuzzy set theory to assign different weights of entity characteristics, which divides entity characteristic weights into local weights and global weights.
   
   (1) Global weight. Different entity characteristics have different effects on entities. According to experience, the weight of data between entities is more than that of entities. The weights of class k features of all entities are expressed in \( W_k = \lambda \), \( \lambda \in (0,1) \).

   \[ W_k = \lambda, \ \lambda \in (0,1) \]  \hspace{1cm} (1)

   Where \( \lambda \) is obtained by experience.

   (2) Local weights. Different characteristics have different effects on a single entity. Let \( W_{ij} \) represent the weight of the entity \( E_i \) of class J characteristics.

   \[ w_{ij} = \frac{N_{ij}}{\sum_{j=1}^{N_i} N_{ij}} \]  \hspace{1cm} (2)

   In which \( N_{ij} \) represents the number of the entity \( E_i \) of class J characteristics. There are four kinds of relations among entities. They are one-way Association, two-way Association, self-correlation and multidimensional association. Therefore, the number of associated relations among entities may be more than 1. Formula (3) is the representation of feature vector \( C_i \) of the entity \( E_i \) after the weight value processing.

   \[ C_i = (W_{i1}^1 \ast C_{i1}, W_{i2}^2 \ast C_{i2}, \ldots, W_{ij}^j \ast C_{ij}); \]

   \[ W_{ij}^j = W_j \ast w_{ij} = \frac{\lambda N_{ij}}{\sum_{j=1}^{k} N_{ij}} \]  \hspace{1cm} (3)

   Among them, \( W_{ij} \) said after the feature vector of local and global weight calculation after the entity \( E_i \) class J feature weights.

2. Similarity calculation method based on information loss

   Information loss is based on information bottleneck theory. Based on information loss similarity calculation method, which is suitable for irregular shape data sets, and is not sensitive to noise points.

   In the process of clustering, the entities with the smallest information loss are clustered, and the characteristic information of the entities after clustering is preserved to the maximum extent. Information bottleneck is based on information theory. The basic idea of information bottleneck is: Let the entity set is \( E=(E_1, E_2, \ldots, E_i) \), whose characteristics set is \( C=(C_1, C_2, \ldots, C_i) \). Because entity association is weak in entity correlation, the mutual information \( I(E_i;E_j) \) of entity \( E_i \) and entity \( E_j \) in the clustering process is minimal. As the same time, the entity’s characteristic information \( C_m \) is retained as much as possible when the entity is merged, so the mutual information \( I(C_m, C) \) of cluster is the largest. This is the information bottleneck of the characteristic \( C_m \) corresponding to the entity \( E_i \) in the clustering process. In order to merge the two entities with the most similarity, two entities with minimum information loss need to be selected. The information loss in the clustering algorithm refers to the difference of mutual information between the entity \( E_i \) and the entity \( E_j \) before and after the merger, and the calculation formula is as follows:

   \[ \sigma(E_i, E_j) = I(E_k, C) - I(E, C) \]  \hspace{1cm} (4)
I (Ek, C) represents the mutual information of entity feature vectors before clustering, and I (E, C) represents the mutual information of the entity's eigenvectors of the merged entity. According to the information theory, the information loss formula is further derived as follows:

\[
\sigma(E_i, E_j) = \sum_{m \in [1,6]} P(E_i) * D_{KL} \left( P(C_m | E_i) \| P(C_m | E_j) \right) + \sum_{m \in [1,6]} P(E_j) * D_{KL} \left( P(C_m | E_j) \| P(C_m | E_i) \right)
\]

\[
D_{KL} \left( P(C_m | E_i) \| P(C_m | E_j) \right) = \sum_{m \in [1,6]} P(C_m | E_i) * \log \frac{P(C_m | E_i)}{P(C_m | E_j)}
\]

In the formula, P(Cm,Ei) represents the probability of the m characteristic of the entity Ei. Ei is a cluster after clustering. The smaller the information loss, the greater the similarity between entities. Based on the information loss level clustering algorithm, the two entities with the least information loss are combined to form entity clusters.

2.2.2 The application of FHCBIIL algorithm in software re-architecting

Based on the two improvements above, the pseudo code of the FHCBIIL algorithm applied to the software re-architecting is as follows:

**Pseudo code of FHCBIIL algorithm**

**Input:** Initial cluster data set \( \{x_1, x_2, \ldots , x_l\} \). The vector of the data object \( X_i \) is \( x_i = (x_i^1, x_i^2, \ldots , x_i^k) \). \( i, k = 1,2,\ldots ,n \). \( x_i^k \) represents the k feature attribute of the i cluster.

**{ Phase I} Data preprocessing**
1. for number of entities \( i = 1,2,\ldots ,n \)
2. for entity characteristic number \( j = 1,2,\ldots ,k \)
3. The eigenvalue expectation of the entity \( E_i \) is \( \bar{v}_i = \frac{1}{n} \sum x_{ij} \); 
4. Eigenvalue variance of entity \( E_i \) is \( \sigma_i = \frac{1}{n} \sum (x_{ij} - \bar{v}_i)^2 \); 
5. The eigenvalues of the entity \( E_i \) after normalization \( x_{ij} = \frac{x_{ij} - \bar{v}_i}{\sigma_i} \);
6. end 
7. end
8. //Building the eigenvector of the entity \( C_i \)
9. for number of entities \( i = 1,2,\ldots ,n \)
10. Eigenvector weights \( W_i = \frac{1}{\sum_{j=1}^{k} |x_{ij}|} \)
11. The eigenvector of the entity \( E_i \) is \( C_i = \left( W_i^1 * C_{i1}, W_i^2 * C_{i2}, \ldots , W_i^k * C_{ik} \right) \);
12. end

**{ Phase II} FHCBIIL algorithm clustering**
12. if number of cluster clusters \( n > n_0 \) & number of cluster layers \( N < N_0 \)
13. repeat 
14. for number of entities \( i,j = 1,2,\ldots ,n \)
15. \( \sigma (\varepsilon_i, \varepsilon_j) \) 
16. Calculating the information loss between two entities according to the formula (5)
17. Find the smallest loss of information \( \sigma (\varepsilon_i, \varepsilon_j) = \min \{ \sigma (\varepsilon_i, \varepsilon_j) \} \);
18. Clustering entities \( \varepsilon_i, \varepsilon_j \), form clusters \( \tilde{B}_i; \)
19. Clusters number self-subtraction \( n - \gamma \);
20. Layer times self-addition \( N + \gamma \);
21. end

**{ Phase III} Generating new tree software architecture**
22. Update dataset \( \tilde{T} \)
23. Update the eigenvector of the clusters \( \tilde{C} = \alpha * C_{i} + \beta C_{j} \);
24. Until \( n \leq n_0 \) & \( N \geq N_0 \)
25. end

**Output:** A new software architecture.
3. Implementation of Software Re-architecting and Evaluation of New Architecture

3.1 Implementation of Software re-architecting Based on FHCBIL Algorithm

In this experiment, we selected the "open experiment management system of University" as a legacy system, which was developed by object-oriented languages. The legacy system source code was converted into source data by reverse engineering tool, which is Enterprise of Architect software. Generating tree graph: the FHCBIL algorithm experiment results are carried out by the data assigned by the 5:5:5:3:1:1 ratio of the global weight value, as in Figure 2:

Figure 2. The new architecture of legacy system based on FHCBIL algorithm

According to figure 2, it can be concluded that the FHCBIL algorithm based on information loss calculation not only achieves a clear level of structure attempt when clustering, but also has less information loss when clusters are merged at the early stage, and only when the clustering is about to end, the information loss is large.

3.2 Implementation of Software re-architecting Based on FHCBIL Algorithm

The accuracy evaluation of the new entity division is reflected by the error rate of the similarity between the entities. The rate of change of similarity between entities is defined as follows:

Let $S_{ij}$ be the similarity of entity $E_i$ and $E_j$ initial clustering, $S_{ij}^*$ clustering similarity at the end, then the similarity of entity $E_i$ and $E_j$ error rate $\rho(E_i, E_j)$ as the calculation formula is as follows:

$$\rho(E_i, E_j) = \frac{|S_{ij} - S_{ij}^*|}{S_{ij}}$$  \hspace{1cm} (6)

According to formula(6), the similarity error rate of all cluster entities before and after clustering is calculated respectively. The entity with the similarity error rate greater than 0.5 is used as the entity, which is wrongly divided in the clustering process, and the entity whose similarity error is less than 0.5 is the entity that is correctly divided. This experiment calculates the number of entities in FHCBIL algorithm, CURE algorithm and DBSCAN algorithm. $N_r$ represents the number of entities that are correctly partitions, and $N_e$ represents the number of entities that are wrongly partitions. The accuracy rate of the entity division $A_c$ is as follows:

$$A_c = \frac{N_r}{N_r + N_e}$$  \hspace{1cm} (7)

The accuracy rate of the three algorithms is shown in table1.
Table 1. Comparison of the accuracy of the three algorithms for entity partitioning.

| Algorithm | Number of Entities | Accuracy Rate |
|-----------|--------------------|---------------|
|           | \( N_r \) | \( N_e \) | \( A_c \) |
| FHCBIL    | 36       | 6           | 0.8571 |
| CURE      | 29       | 13          | 0.6905 |
| DBSCAN    | 34       | 8           | 0.8095 |

Through comparison and analysis of table 1, we can draw a conclusion that the number of correct classification FHCBIL algorithm, CURE algorithm, BDSCAN algorithm is higher than error division. The exact number of entities divided and the accuracy rate of entity partition by FHCBIL algorithm are higher than that of the CURE algorithm and the DBSCAN algorithm. This shows the high quality of the new architecture after software re-architecting based on the FHCBIL algorithm.

4. Concluding Remarks
In this paper, the clustering idea is applied to the software re-architecting, a hierarchical clustering algorithm is proposed based on the information bottleneck (FHCBIL). This algorithm extends the feature weights, and uses the loss of information as similarity calculation. FHCBIL algorithm solves the problem that the relationship between two elements, which can not reflect the influence degree of entity characteristics to entities. FHCBIL algorithm divides the weight value into the global weight and the local weight that fully consider the effect of the feature on the entity. Experiments show that the application of FHCBIL algorithm in the software re-architecting of legacy system not only achieves a new architecture with high cohesion and low coupling, but also has high accuracy of entity division.

5. References
[1] Onaiza Maqbool, Haroon A. Babri. Hierarchical Clustering for Software Architecture Recovery [J]. 2007, 33(11):759-780.
[2] Masoud Hamid, Jalili Saeed. A clustering-based model for class responsibility assignment problem in object-oriented analysis. Journal of System and Soteware. 2014,93:110-131.
[3] Mkaouer Mohamed Wiem, Kessentini Marouane, Bechikh Slim, Cinneide Mel O, Deb Kalyanmoy. On the use of many quality attributes for software refactoring: a many-objective search-based software engineering approach. Empirical Software Engineering. 2016,21(6):2503-2545.
[4] Software Research; New Findings from COMSATS University Describe Advances in Software Research (DOOR: Distributed Object Oriented Software Restructuring Approach Using Neural Network)[J]. Computer Technology Journal, 2020.
[5] Han Li. Research on Several Issues of Legacy System Reengineering[D].Dalian University of Technology,2013.
[6] Zhang Xinyue. Software Sustainable Development Progress and Evaluation and its Application in Architecture Reconstruction [D]. Southeast University, 2018.