A Data Mining Method on Software System Layer

Liu Yonghui
Computer and Information Engineering Dept.
Baoding Vocational and Technical College
Baoding, China,13930828561
Lyhui@163.com

Wang Yanhui
Computer and Information Engineering Dept.
Baoding Vocational and Technical College
Baoding, China,15354328896
Wangyh@126.com

Hu Yi'nan
Modern Education Dept.
Baoding Vocational and Technical College
Baoding, China,13582398421
HUYINAN@126.com

***Abstract*—This paper puts forward a data mining method on software system layer which is based on researches on clustering data mining. At first, this method gathers all kinds of software data. Then it extracts characters from them to mark softwares on Halstead software science. Thirdly, it classifies the softwares in different categories. For the softwares in the same category, it regards them as the same software cost or the similar software structure. For the softwares in the different category, it finds the essential different factors among them. At last, we show 5414 experiment results in software system data mining. So we can conclude that this method is feasible and effective.**

Keywords- data mining; software system layer; clustering

I. INTRODUCTION

The software engineering has a important status in development and application of computer. But with the development in size and complexity, some problems troubles the software engineering more and more. The problems includes: "how to compare softwares", "what is the critria for the classification of software modules", and etc.. For example, if we make no plan or implement blindly when program, a result can be make that the budget is broken through. If we don't understand the customers' needs, fully, we shall meet some problems in the late stage of programming. If we don't program specificatedly or have no complete information about program, the result can also be failure. Such problems we always meet in the process of programming.

After the conception of fuzzy sets was bring forward by L. A. Zadeh in 1965, the clustering method has more and more application fields. The classical clustering theory is as the follow. If the similarity of A and B is more than a specified threshold and so to B and C or A and C, A, B and C belong to the same cluster. That is to say that the quality of clustering is determined by the formula for calculating similarity. In fact, we get the formula by extracting feathers from practical problems. We need the feathers without redundancy and distortion. Because of the advantages of the clustering method, it was applicated in data mining.

II. APPLICATION OF CLUSTERING IN DATA MINING

A. *Classification method*

Classification method is a common one in clustering methods. It classifies a special set into k subsets according to a certain principle of classification. Then it processes data in k-means method or in k-medoids method.

B. *Model method*

Model method constructs a data model at first. Then it fits the given data with the model. At the end of clustering, each clustering is regarded as a new feature set.

C. *Density method*

Density method is a clustering based on distance between objects. At first, it sets a certain density threshold. Then it obtains clusters constantly until the density close to reference point is less than the certain threshold.

D. *Fuzzy clustering method*

Fuzzy clustering method leads fuzzy mathematics into clustering. Its character is to soften the boundary of clusters.At first, it gives a value to indicate the similarity between two objects. And the value is between 0 and 1. Then it gets a matrix composed with similarity coeffcients. After solving the transitive closure operator, it gets a fuzzy equivalence matrix. At last, it outputs the results of clustering.

III. FEASIBILITY OF CLUSTERING IN DATA MINING ON SOFTWARE SYSTEM LAYER

A. *A brief description of system*

Assuming a project W has n project files (p_1, p_2, p_3, $\cdots p_n$) and each project files has t different modules (m_1, m_2, m_3, $\cdots m_t$), we can construct the project from bottom to top. We evaluate the project from modules. If we reguards each modules as a item set, some links and disciplines exist between different item sets. The task of data mining is to find them.

B. *Process of system*

Firstly, we must mark the modules correctly for a certain project. According to Halstead software science and part

repetitive theory, effectiveness of marking depends on the extracted characters. We choose lines of code, number of different operatings, number of different operators, total number of different operatings and total number of different operators to scale a module object. Then we cluster and simplify the module according to the similarity of modules in a same project. At last we can get a evaluating module of the whole project.

IV. CONCEPTION AND METHOD OF DATA MINING ON SOFTWARE SYSTEM LAYER

A. Definitions

Definition 1: Center Knowledge $\left\{ \vec{s}_1 \quad \vec{s}_2 \quad \cdots \quad \vec{s}_n \right\}$

$\vec{s}_1 \quad \vec{s}_2 \quad \cdots \quad \vec{s}_n$ is n five dimensional vectors.

Definition 2: Accuracy of Knowledge. It is a decimal between 0 and 1 to measure the credibility among different classification.

Definition 3: Floating Domain of Knowledge. It refers to all possible classifications of a group of objects.

Definition 4: Degree of Deviation.It is a scale of degree of optimization of software. It can be get from the following formula.

$$\sigma = \left| N - \left(\eta_1 \log_2^{\eta} 1 + \eta_2 \log_2^{\eta} 2 \right) \right| \tag{1}$$

N is the number of lines of code in program. η_1 is the number of operators in program. η_2 is the number of operand.

B. Process

1) Process of generating center knowledge

Step 1: Get data set of software. Each record takes a line in software data sets and has five fields. The five fields represend lines of code, number of different operatings, number of different operators, total number of different operatings and total number of different operators respectively. If we reguard each record as a vector, we can get five vectors or one five-dimensional vector. So we can describe a program with a five-dimensional vector. The relation between programs can also be calibrated by quantitative relationship between vectors. Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

Step 2: Sampling. Because the scale of software data set is very large, we extract a sample set composed by data choosing from software data set at random.

Step 3: Normalization. We use the following formula to normalize sample data.

$$r = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

x is the value of a certain kind of data. $x_{\min}$ is the minimum of this kind of data. $x_{\max}$ is the maximum of this kind data. We normalize the lines of code, number of different operatings, number of different operators, total number of different operatings and total number of different operators respectively.

Step 4: Getting the fuzzy relation. If there are two vectors i and j, the fuzzy relation between them is as the following.

$$r_{i,j} = 1 - \sqrt{\frac{1}{m} \sum_{k=1}^{m} (x_{jk} - x_{ik})^2} \tag{3}$$

m is the number of dimension of the two vectors. x_{jk} is the normalized value of No.k sub-vector of vector j. x_{ik} is that of vector i.

Step 5: Getting the center knowledge. We can compose a similar relation matrix with fuzzy relation of n vectors. Obviously, it is a symmetrical square. Then we can classify all the vectors with ƛ, a horizonal cut set. The vectors, which is classified in the same category, have the same software cost and complexity.

2) Process of generating knowledge floating domain

Step 1: Calculating the degree of deviation with the following formula.

$$\sigma = \left| N - \left(\eta_1 \log_2^{\eta} 1 + \eta_2 \log_2^{\eta} 2 \right) \right| \tag{4}$$

N is the number of lines of codes. η_1 is the number of different operators. η_2 is that of different operatings.

Step 2: Calculating the similar relation with the following formula.

$$r_{i,j} = 1 - \sqrt{\frac{1}{m} \sum_{k=1}^{m} (x_{jk} - x_{ik})^2} \tag{5}$$

In this formula, $w_{ij} = 1 + Q \frac{(\sigma_i - \sigma_j)}{\sqrt{\sigma_i^2 + \sigma_j^2 - \sigma_i \sigma_j}}$, 0<Q<1, $-1 \le \frac{(\sigma_i - \sigma_j)}{\sqrt{\sigma_i^2 + \sigma_j^2 - \sigma_i \sigma_j}} \le 1$. i and j are two vectors. Deviation of them are $\sigma_i and \sigma_j$.

Step 3: Calculating the floating domain. When Q has different values, we can get different classified sets.

Step 4: Calculating the accuracy of knowledge with the following formula.

$$C = \frac{\left| f(\vec{s}) \cap E(\vec{s}) \right|}{\left| f(\vec{s}) \right|} \ge \min\_correctness \tag{6}$$

a) $f(\vec{s})$ *is classification of center knowledge.* $E(\vec{s})$ *is that of floating domain.*

V. RESULTS OF EXPERIMENT

We make experiment with Visual C++ 6.0. The test data comes from commercial software packages of Borland and Microsoft.

We make the experiment in 5 steps: preprocessing, sampling, center knowledge mining, generating "flaoting domain" and "accuracy" and repeating step 2 to step 4.

After 5 samplings, we get results as the following 5 tables.

TABLE 1 RANDOM SAMPLING RESULTS FOR THE 1ST TIME (N=201, λ=0.8)

Empirical coefficient Q	0.6	0.6	0.9	0.98
Center of knowledge	Category 6	Category 6	Category 6	Category 6
Floating domain	Category 6	Category 6	Category 6	Category 6
Min_correctness	0.5	0.5	0.5	0.5
The actual accuracy C	(0.60,0.65,1,1,1,0.5)	(0.60,0.68,0.93,1,1,0.5)	(0.60,0.68,0.93,1,1,0.5)	(0.60,0.68,0.93,1,1,0.5)

TABLE 2 Random sampling results for the 2ND time (N=158,λ=0.8)

Empirical coefficient Q	0.6	0.6	0.9	0.98
Center of knowledge	Category 6	Category 6	Category 6	Category 6
Floating domain	Category 4	Category 4	Category 4	Category 4
Min_correctness	0.66	0.66	0.66	0.66
The actual accuracy C	(0.87,0.87,0.67,1,1,1)	(0.87,0.87,0.67,1,1,1)	(0.87,0.87,0.67,1,1,1)	(0.87,0.87,0.67,1,1,1)

TABLE 3 RANDOM SAMPLING RESULTS FOR THE 3RD TIME (N=99, λ=0.8)

Empirical coefficient Q	0.6	0.6	0.9	0.98
Center of knowledge	Category 4	Category 4	Category 4	Category 4
Floating domain	Category 7	Category 7	Category 7	Category 7
Min_correctness	0.6	0.6	0.6	0.6
The actual accuracy C	(0.61,1,0.85,1)	(0.61,1,0.85,1)	(0.61,1,0.85,1)	(0.61,1,0.85,1)

TABLE 4 RANDOM SAMPLING RESULTS FOR THE 4TH TIME (N=51, λ=0.8)

Empirical coefficient Q	0.6	0.6	0.9	0.98
Center of knowledge	Category 7	Category 7	Category 7	Category 7
Floating domain	Category 3	Category 3	Category 3	Category 3
Min_correctness	0.7	0.7	0.7	0.7
The actual accuracy C	(0.52,0.52,1,1,1,1)	(0.52,0.52,0.93,1,1,1)	(0.52,0.68,0.93,1,1,0.5)	(0.52,0.68,0.93,1,1,0.5)

TABLE 5 RANDOM SAMPLING RESULTS FOR THE 5TH TIME (N=28, λ=0.8)

Empirical coefficient Q	0.6	0.6	0.9	0.98
Center of knowledge	Category 6	Category 6	Category 6	Category 6
Floating domain	Category 6	Category 6	Category 6	Category 6
Min_correctness	0.5	0.5	0.5	0.5
The actual accuracy C	(0.49,0.49,1,1,1,0.5)	(0.49,0.68,0.83,1,1,0.5)	(0.60,0.49,0.83,1,1,0.5)	(0.60,0.49,0.93,1,1,0.5)

In the tables above, n is the number of samples. λ is the level of cut set of clustering.

VI. CONCLUSION

This paper tries to proposed a method in which data mining can be used in software layer. Firstly, we collect different software systems. Then we mark them with feathers extracting from them. Lastly, cluster them in the clustering method. Data mining is a hot and difficult research areas in software engineering and science.

This paper also tries to applicate data mining in software engineering. And the simulation experiments show that the data mining method on software system layer is efficient in prediction of software cost and evaluation of software complexity.

REFERENCES

[1] LI Ji,LI Xiao-ming,LU Sang-lu,CHEN Gui-hai,XIE Li(State Key Laboratory for Novel Software Technology,Nanjing University,Nanjing 210093,China);Applying Sieving Technique in Parallel I/O[J];Acta Electronica Sinica;2001-02

[2] WANG Xiao-han,CHEN Jie (Key Laboratory of Intelligent Computing and Signal Processing,Anhui University,Hefei,230039,China;Department of Computer Science and Engineering,Anhui Normal University,Wuhu,241000,China);The Design & Implementation of Fuel Management Intelligent Analytical Network System[J];Journal of Anhui Institute of Education;2006-03

[3] ZHOU Hua,MI Hao (School of Education,Anqing Teachers College,Anqing 246133,China);Using Alternative Covering Design Algorithm to Valuate the Reputation of Life Insurance

Customer[J];Journal of Anqing Teachers College(Natural Science Edition);2009-02

[4] HAN Min~*,CUI Pi-suo(School of Electr.and Inf.Eng.,Dalian Univ.of Technol.,Dalian 116024,China);A dynamic RBF neural network algorithm used in pattern recognition[J];Journal of Dalian University of Technology;2006-05

[5] ZHANG Cheng-gong,HUANG Di-ming,HU De-kun (School of Computer Science and Engineering, Univ.of Electron. Sci. & Tech. of China Chengdu 610054);An Anti-Spam System AIASS Based on Artificial Immune Principle[J];Journal of University of Electronic Science and Technology of China;2007-01

[6] WANG Bin (Modern Education Technology Center,Guangdong College of Industry and Commerce,Guangzhou 510510,China);Recognition Method for External IM User Based on Secondary Mining[J];Computer Engineering;2010-22

[7] TANG Jin-Tao,WANG Ting,WANG Ji(College of Computer,National University of Defense Technology,Changsha 410073,China);Shortest Path Approximate Algorithm for Complex Network Analysis[J];Journal of Software;2011-10

[8] CAO Bang-Hua HU Hong-Jun ZHANG Da-Peng ZHU Xiao-Da(Forestry College of Shandong Agricultural University,Taian Shandong 271018,China);Rooting Capacity and Correlative Enzymes Activities of Hardwood Cuttings of Mulberry[J];Science of Sericulture;2008-01

[9] Ou Wenlin(Chimen Forestry Station of Yanping District Forestry Bureau in Fujain Province Nanping353000);Analysis on Effect of Low-Yield Forest of Phyllostachys pubescens Mazel[J];Hubei Forestry Science and Technology;2008-02

[10] He Yong Chen Shiping College of Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;DESIGN AND IMPLEMENTATION OF DATA SHARE IN CAMPUS BASED ON WEB SERVICE[A];[C];2005