A Study on Applying KFCM Algorithm to Source Code Mining

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

Tian Jingjing
Computer and Information Engineering Dept.
Baoding Vocational and Technical College
Baoding,China,15354328896
Tianjing@126.com

Zhang Lei
Modern Education Dept.
Baoding Vocational and Technical College
Baoding,China, 15354328896
zhanglei@163.com

***Abstract*—This paper provides a algorithm, which is based on that kernelized fuzzy C-means uses on the study of source code mining, to solve the problem that the large number of quantities, multiple attributes and most of them discrete of software engineering. By using this algorithm, we can improve the efficiency of mining and seek faster and more effective cluster approaches. Meanwhile, we can also solve the problem that the KFCM algorithm can not cluster text data directly. Then we can over the defect of only being able to obtain the minimum values by integrating KFCM and genetic algorithm. Finally, the experiment shows that the improved KFCM algorithm has a good clustering performance and high efficiency on data mining.**

Keywords-C-means;KFCMalgorithm;source code mining

I. INTRODUCTION (HEADING 1)

Source code mining is becoming more and more concerned with the expansion of the research areas of data mining. KFCM is used in the research of source code mining to find the dicipline hiden in source code. In this paper, we combine KFCM with genetic algorithm to get a global optimal solution.

II. DATA PREPROCESSING

We preprocess data in the way of TF-IDF, because discrete text data can not be clustered by KFCM.The weights of the feature are represented by the product of term frequency and inverse document frequency.

$$W = TF \times IDF = TF \times \frac{1}{DF} \tag{1}$$

In the above formular, TF is the occurrence frequency of entry (t) in document (d). And IDF is the inversely proportional to the occurrence frequency of entry (t) in document.

The basic idea of this theory is that if the quantity of document include entry t is smaller and the value of IDF is bigger, entry t have a good category distinguishing ability. On the contrary, t have a poor one. In fact, if a entry appears in a type of documents frequently, it can represent the characteristics of this type of documents appropriately. And we can make it as a feature word of this type of documents to distinguish to another types.

III. KERNELIZED FUZZY C-MEANS

We assume X is a data sample collection and $X = \{x_1, \ x_2, \ \cdots, \ x_n\}$. Firstly, we transform input space (χ) transform into a high-dimensional feature space (F) by a nonlinear mapping, $\phi : x \to F(X \in R^n \to \phi(x) \in R^p$. Then we cluster in feature space. In this way, the objective function of KFCM in feature space is as the following.

$$J_m = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \left\| \Phi(x_k) - \Phi(v_i) \right\|^2 \tag{2}$$

In this formular, V is clustering center and $V = \{v_1, \ v_2, \ \cdots, \ v_n\}$. $\phi(v_i)$ is image of clustering center in corresponding feature space. U is fuzzy partition matrix and $U = \{u_{ik} \in [0,1]\}_{c \times n}$. u_{ik} is the membership of No.k data to No. i type and $\sum_{i=1}^{c} u_{ik} = 1$, $\forall k$, $0 \prec \sum_{k=1}^{n} u_{ik} \prec 1$. m is fuzzy control index and it controls the degree of fuzzy. We use $\left\| \phi(x_k) - \phi(x_i) \right\|^2$ to represent the distance between center v_i and original data (x_k) mapped by ϕ in feature space.

In this paper, we use Gaussian function as Kernel function.

$$K_G(x, y) = \exp\left(-\frac{\left\| x - y \right\|^2}{2\sigma^2} \right) \tag{3}$$

The step of KFCM is as the following:

Step 1: Update the membership by the following formular. Then we can get the partition matrix $U^{(b)}$.

$$u_{ik}^{(b)} \frac{(1/K(x_k,x_k)+K(v_i^{(b)},v_i^{(b)})-2K(x_k,v_i^{(b)}))^{1/(m-1)}}{\sum_{j=1}^{c}(1/K(x_k,x_k)+K(v_i^{(b)},v_i^{(b)})-2K(x_k,v_i^{(b)}))^{1/(m-1)}} \quad (4)$$

Step 2: Update the clustering centers by the following formular. Then we can get the clustering centers matrix $V^{(b+1)}$.

$$v_i^{(b+1)} = \frac{\sum_{k=1}^{n}(u_{ik}^{(b)})^m \tilde{K}(x_k,v_i^{(b)})x_k,}{\sum_{k=1}^{n}(u_{ik}^{(b)})^m \tilde{K}(v_i^{(b)},x_k)} \quad (5)$$

Step 3: If $\left\|V^{(b)}-V^{(b+1)}\right\| \prec \varepsilon$, the algorithm stops, then outputs the partition matrix U and clustering centers matrix V. Otherwise, makes b add 1, then turn back to Step 1.

The algorithm clusters in feature space (R^y). KFCM is only a algorithm which has a good ability in local search. And it always gets a local optimal solution but a global one.

IV. KFCM BASED ON GENETIC ALGORITHM

Because of the above disadvantages of KFCM, we use genetic algorithm to improve it.

A. Coding and population initialization

The substance of real number coding, which makes clustering center as chromosome, is that look the string composed by c clustering centers as a chromosome. Additionally, each clustering center has r properties. So the length of chromosome is a code-string of real number ($c \times r$). Initializing population, we can get Initialized population composed by M chromosomes.

Step 1: Generate the partition matrix (U) at random.

Step 2: Calculate the initial clustering center matrix (V) according to formular (5).

Step 3: Make the clustering centers as a chromosome according to encoding scheme.

B. Fitness function

The fitness function used in this paper can be defined as the following.

$$Fitness = 1/\left(1+\sum_{1=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}\left\|\Phi(x_k)-\Phi(v_i)\right\|^2\right) \quad (6)$$

C. Genetic operation and optimization of KFCM

Selected operators are used to realize survival of the fittest of unit in group in genetic algorithm. Because KFCM has a good ability in local search, the results must be KFCM optimized after a running of genetic algorithm. The optimized population comes into the next generation group in order to improve the convergence speed.

D. Clustering algorithm

The process of KFCM based on genetic algorithm is as the following.

Step 1: Initialization. Input sample X, the number of clustering c, the scale of group M, crossover probability Pc, mutation probability Pm and the maximum generation maxgen.

Step 2: Coding. Initialize the group at random and get P(0).

Step 3: for t=1:maxgen. Calculate fitness of the unit in P(t) group. Get the next generation group P(t+1).

Step 4: Get the optimal unit in the last generation after cycle of operation.

V. EXPERIMENT DESIGN AND RESULTS

We use FCM, KFCM and SGAKFCM in the experiment with Weka3.5.71 to test the clustering effect and efficiency of the three algorithm. We choose software of Matlab 2008a and hardware of Pentium 4 3.0 GHz and Memory 1G.

A. Built data input model

We choose Class and ClassMember to built model. The input model is as the following table 1 and 2.

TABLE 1 CONSTRUCTION OF THE CLASS INPUT MODEL

Property	ClassID	ClassName	IsInherits	InheritsFrom	Package Name	IsImplement	Implement To
Description	ClassID	ClassName	ClassID	SuperClassName	Package Name	Yes/No	Implement Name

TABLE 2 CONSTRUCTION OF THE CLASSMEMBER INPUT MODEL

Property	MemberID	MemberName	Type	ClassName	IsStatic	IsBasic Type	Modifier
Description	ClassMemberID	ClassMemberName	ClassMember Type	ClassName	Yes/No	Yes/No	Sysymbol

B. Experiment result

We make experiment with 994 classes and 4116 classmembers of Weka code. The results are as the following table 3 and 4.

TABLE 3 CLUSTERING RESULTS IN THE DATA SET ON THE 994 CLASS MEMBERS

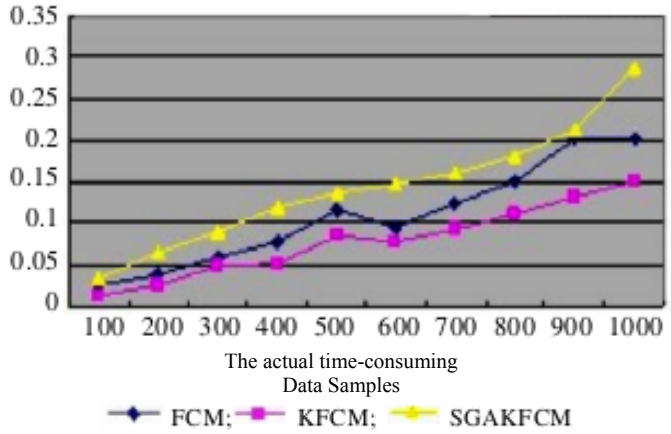
Class			
Algorithm	The Amount Of Data	Optimal Number Of Clusters	The Actual Number Of Clusters
FCM	994	13	9
KFCM	994	12	10
SGAKFCM	994	12	12

TABLE 4 CLUSTERING RESULTS IN THE DATA SET ON THE 4116 CLASS MEMBERS

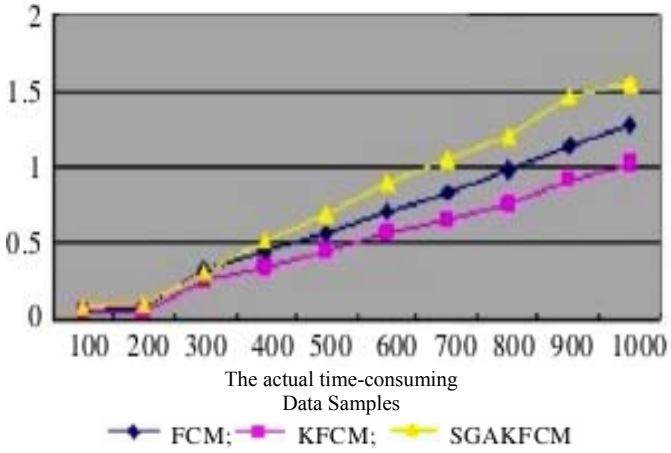
Class Member			
Algorithm	The Amount Of Data	Optimal Number Of Clusters	The Actual Number Of Clusters

FCM	4116	20	19
KFCM	4116	25	23
SGAKFCM	4116	24	23

To compare the time complexity of 944 classes of the three algorithm, we can get the efficiency of them as the following fig.1 when c=5, 15.

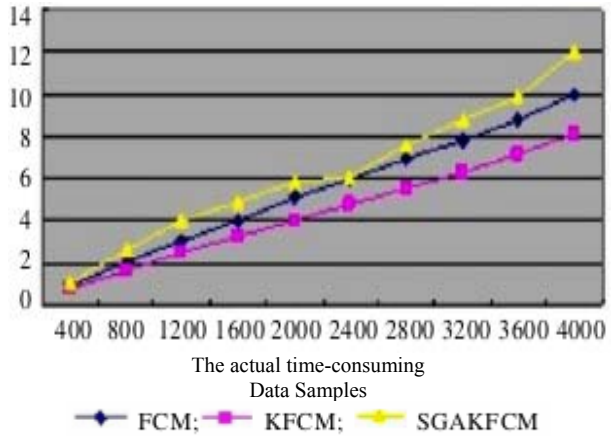


(a) c=5 Efficiency in the implementation

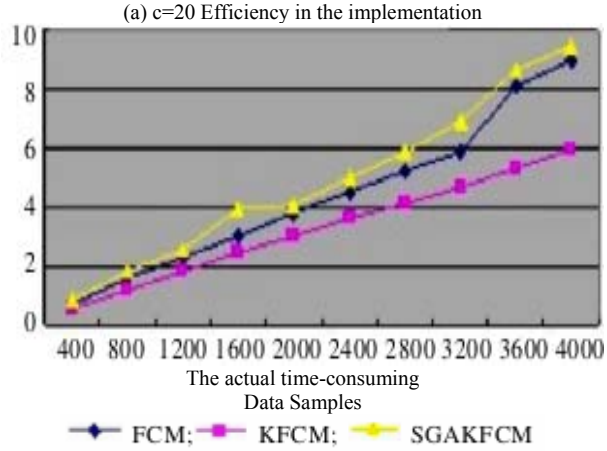


(b) c=15 Efficiency in the implementation

Figure 1 C for different valuesof the three algorithms on the efficiency of the implementation of the 994 class Otherwise, to compare the time complexity of 4116 class members of the three algorithm, we can get the efficiency of them as the following fig.2 when c=15, 20.



(a) c=20 Efficiency in the implementation



(b) c=15 Efficiency in the implementation

Figure 2 C for different valuesof the three algorithms on the efficiency of the implementation of the 4116 class

VI. CONCLUSION

In this paper, we use TF-IDF to remedy the defect of that KFCM can not process discrete text data directly. Meanwhile, we ues SGAKFCM in code data mining. And the results of experiment show that the operating efficiency of KFCM is significantly higher than that of FCM. And KFCM improved by genetic algorithm solves the problem of local optimization of KFCM.

REFERENCES

[1] XU Xiu-hua,CHENG Xiao-jin,LI Ye-li(Beijing Institute of Graphic Communication,Beijing 102600,China);Research on Decision Support System of Publishing Based on Data Warehouse[J];Journal of Beijing Institute of Graphic Communication;2008-04.

[2] XU Zhi-wei(Computer Science and Technology Institute,Changchun University,Changchun 130022,China);An analysis on common problems of using scanf() function in C language[J];Journal of Changchun University;2008-12

[3] KONG Pan1,DENG Hui-wen2,JIANG Huan1,HUANG Yan-yan1(1.School of Computer and Information Science,Southwest China University,Chongqing 400715,China;2.Institute of Logic and Intelligence,Southwest China University,Chongqing 400715,China);Improved kernel-based fuzzy clustering algorithm[J];Journal of Computer Applications;2008-09

[4] Yuan Yunneng Wu YangCheng Gong(School of Electronics and Information Engineering,Beijing University of Aeronautics and Astronautics,Beijing 100083,China);Simplified method of kernel fuzzy c-means clustering for image texture classification[J];Journal of Beijing University of Aeronautics and Astronautics;2008-03

[5] YANG Hua-fen(Dept.of Computer Science,Qujing NormalCollege,Qujing 655000,China);Short-term load forecasting in power system based on improved fuzzy neuro net[J];Journal of Changchun Institute of Technology(Natural Sciences Edition);2009-01

[6] Li Dandan (School of Business,Tianjin University of Finance and Economics,300222);An E-commerce Recommendation Method Based on Genetic Fuzzy Clustering[A];[C];2008

[7] ZHANG Yufang1,PENG Shiming1,LV Jia2(1.Department of Computer Science,Chongqing University,Chongqing 400045;2.College of Mathematics and Computer Science,Chongqing Normal University,Chongqing 400047);Improvement and Application of TFIDF Method Based on Text Classification[J];Computer Engineering;2006-19