I.J. Education and Management Engineering **2012, 5, 70-79**
Published Online May 2012 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2012.05.12

Available online at http://www.mecs-press.net/ijeme

Application Research on Data Mining Methods in Information Communication Mode of Software Development

Caixian ye[a], Gang Zhang[b]

[a]*IT Department, NiuTaiLai communication equipment Co.Ltd., GuangZhou , China*
[b]*Faculty of Automation, Guang Dong University of Technology, GuangZhou, China*

Abstract

Smaller time loss and smoother information communication mode is the urgent pursuit of the software R&D enterprise. Information communication is difficult to control and manage and it needs more technical to support. Data mining is an intelligent way tried to analyze knowledge and laws which hidden in massive amounts of data. Data mining technology together with share repositories can improve the intelligent degree of information communication mode. In this paper, the framework of intelligent information communication mode which based on data mining technology and share repositories is advanced, and data mining model for information communication of software development is designed. In view of the extant single decision tree algorithm existence the characteristics that counting inefficient and its learning based on supervise, a new semi-supervised learning algorithm three decision trees voting classification algorithm based on tri-training (TTVA) is proposed. This algorithm in training only requests a few labeled data, and can use massively unlabeled data repeatedly revision to the classifier. It has overcome the single decision tree algorithm shortcoming. Experiments on the real communicated data sets of software developmental item indicate that TTVA has the good identification and accuracy to the crux issues mining, and can apply to the decision analysis of the development and management of the software project. At the same time, TTVA can effectively exploit the massively unlabeled data to enhance the learning performance.

Index Terms: Information communication mode of software development; data mining; share repositories; semi-supervised learning; Three decision trees voting classification algorithm based on Tri-training (TTVA)

1. Introduction

Software development is a team's activities with specific goals. Only through effective communication the team is in order to play their collective strengths. Information communication includes all the information to send, to receive, to transfer, and the interaction of this process [1, 10]. In software development, all to obtain

* Corresponding author.
E-mail address: [a]yecaixian@tom.com, [b]ipx@gdut.edu.cn

information from outside their own activities belong to the scope of the information communication. Information communication data have some characteristics such as reconcilability, transitivity, can be transformative and can be stored and others, these characteristics create the conditions for the information collection and reuse and afford idea for the intelligent information communication mode to apply data mining technologies and sharing technologies.

Data mining is the analysis of (often large) observational data sets from the database, data warehouse or other large repository incomplete, noisy, ambiguous, the practical application of random data to find unsuspected relationships and summarize the data that are both understandable and useful to the data owner. It is a means that data extraction, cleaning and transformation, analysis, and other treatment models, and automatically discovers the patterns and interesting knowledge hidden in large amounts of data, this helps us make decisions based on a wealth of data. Information communication mode of software development lies in how to collection, analysis, and mine out the hidden useful information in the various data from information communication between developers and the staff interaction with manages, and then used the knowledge to make decision.

In the use of networked environments, modern information technology making the information generated and communicated has undergone a qualitative leap. Information communication mode has been preliminary discussion [2, 11], but it has not been a mature intelligent mode, which consider the application of repository and data mining technology very few. Contemporary software project management theories have lack of adequate concerns of the information communication of development team. A survey shows 40 hours one week of development engineers, the actual development time only 16 to 18 hours average, and the most other time consumed in the personnel communicates among developers [10]. Information communication has a profound impact on software development. Project managers and developers need a good information communication mode to communicate, information communication mode requires good technical to support intelligence, and intelligent information communication mode has some positive effects, not only to help project managers to make informed decisions, provide the right solution for technical staff who encountered the problems, but also to help the entire development team communicate with each other, consistent, saving time and so on. However, traditional information communication mode common practice to use internal communications, such as staff training, technical meetings, or members discussed and others, this mode lack of tool support, low efficiency, loss Tai. From the actual need for the development team, start from intelligent information communication mode, using data mining methods to mine communicated information on the large historical data, and with data mining technology to find the real knowledge for share repositories. The purpose is to improve the efficiency of information communication, increase knowledge sharing, to make informed decisions by managers.

The rest of this paper is organized as follows: Section 2 describes the basic concepts of intelligent information communication mode of software development. Section 3 presents the information communication mode based on data mining, Part 4 describe the experimental results and discussion; Part 5 concludes and raise several issues for future work.

2. Information communication mode of software development

Apply a combination technology of data mining and share repositories to carry out intelligent information communication mode. Software development wants to minimize time loss is to reduce communication overhead that people directly involved in activities, to reduce the cost of information exchange is to improve the degree of sharing information.

A. The framework of Intelligent Information communication mode

Through data mining and share repositories, the programmer need not directly encounter problems in consultation with other staff, but obtained currently the best solution from the classified share repositories. The high-level leaders and internal review members can get the data from the teen members to analysis with project leaders. Based on historical statistical data to assess the characteristics and background of communication problems, to prevent relation problems may result in the loss. Therefore reducing the cost of human communication, maintain an active information communication mode, and the performance of mining information for share repositories by the use of data mining can improve sharing capability. The framework of intelligent information communication mode shows in Figure 1.

Note:

a) Communicated data sets: the experience through the development team members, the actual problems encountered, the expert information and the internet resources are collected to establish a data warehouse or data marts.

b) Share repositories: The technical staffs preprocess and mine on communicated data sets to get the knowledge and solutions to problems, then tapped into the classified share repositories. Managers can make informed decisions and the technical staff of the problems encountered can get a viable solution from share repositories.

c) Data mining methods set: data mining algorithms and modes designed by the technical staff.

d) Arrow and its direction: the direction and line of information transmission.

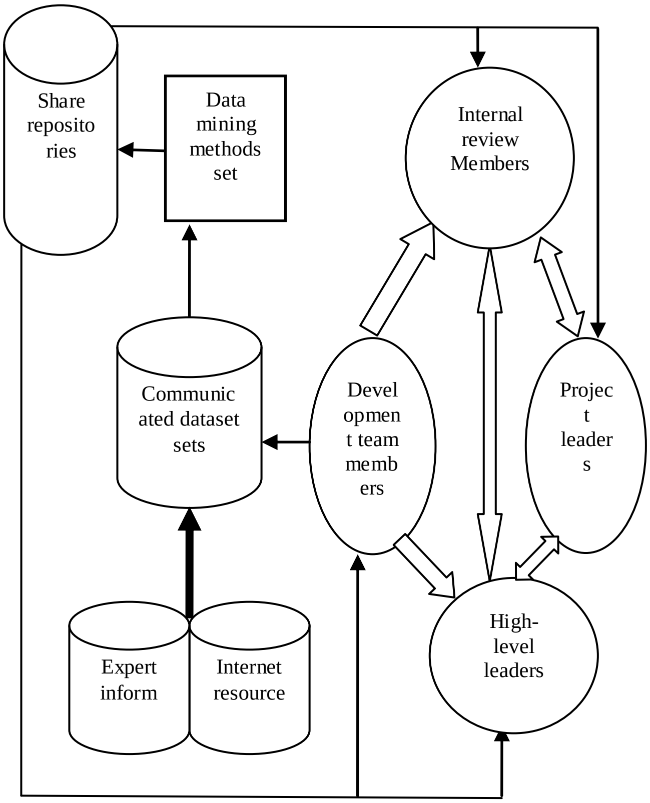

Fig 1.The framework of Intelligent Information communication mode

B. Characteristics of Intelligent Information communication Mode

a) Application of data mining technology. Use of data mining methods approach to the communicated data to mine knowledge, then knowledge will respective tap into the classified share repositories. Or the optimal priority solution to the communicated question will mine from the share repositories. And identification and estimation of types and causes of the communicated data, problems and errors caused by the induction of factors, so that can effectively control and reduce the incidence of major mistakes.

b) The classification share repositories are required. Share repositories to improve the speed of information transmission and reduce the frequency of direct communication degree between people. Every one can acquire knowledge through the share repositories.

c) Strong communication mode has natural time loss. Share repositories make smooth communication channels for the team members, there is no any direct communication, direct communication is necessary only when share repositories can not provide this information.

3. The analysis of information communication mode based on data mining

Apply data mining technology to intelligent information communication mode will play an increasingly important role. Machine intelligence can replace human communication, education and training, it can provide immediate while not constraints by time and space, and can work in parallel, thus greatly reducing the losses caused by direct interaction.

C. The data mining model of information communication

Technical staff pre-processed and mined through the communicated data sets to discover useful knowledge. Data pre-processing [6, 13] refers to pre-process the large amounts of the noise data which is not complete and inconsistent, generally including data summarized, data transformation, data integration and data cleaning. Data cleaning can be used to remove noise in the data, fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies; data integration merged multiple sources into the same data storage that is integration of multiple databases, data cubes, or files; data transformation for normalization of data; data reduction can reduce the data size by gathering, removing redundant features or clustering methods. The design of mining algorithms and models are the data mining module main work. The data mining model of information communication shown in Figure 2:

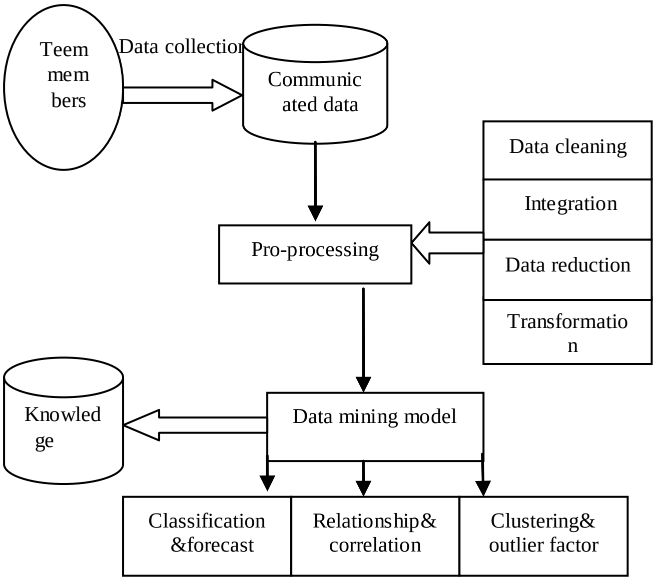

Fig 2. Data mining model of information communication

D. TTVA apply to the information communication mode

Some crux issues appear in the software development which bring the communication frequently, the development interrupt, or data missing and so on, these situations often cause a very big loss. Moreover the crux issues were not easily discovered, if these issues cannot distinguish promptly, they may bring the very big trouble to the development. In the communication, the characteristics that divergence and two types property (Y/N) of the issues suit to use the decision tree classification. Because the issue data set majority is unlabeled data, proposed three decision trees voting classification algorithm based on Tri-training (TTVA) to obtain the high accuracy classified data.

1) Analysis TTVA

The essence of semi-supervised learning algorithms is to enhance accuracy of certain correlation statistical distribution estimate through massive unlabeled sample. Zhou [7] has proposed a new semi-supervised learning algorithm named tri-training. This algorithm generates three classifiers from the original labeled example set. In detail, in each round of tri-training, an unlabeled example is labeled for a classifier if the other two classifiers agree on the labeling, under certain conditions.

The application of tri-Training algorithms is need to select three initial classifiers. It is know from the integrated strategy that the more difference between the classifiers, the better are the integrated final results. Most of the data properties is unlabeled of the practical applications, in order to make the initial classifiers has certain diversity and adapt to different types of data and also can be increased independence among the three classifiers. That the decision trees classifiers choice three different splitting properties as the separation different eigenfunctions, and this is the greatest degree to avoid tri-training to the three self-training classifiers integrated.

The most important of decision tree classifier is to choose the splitting property. This paper uses three kind of splitting properties --the information gain, the Gini index and based on the Goodman-kruskal synthetic index as these three classifiers' eigenfunctions.

a) information gains

Let D, the data partition, be a training set of class-labeled tuples. Suppose the class label attribute has M distinct values defining M distinct classes Ci (i=1, ...m). let Ci,D be the set of tuples of class Ci in D. let |D| and | Ci,D | denote the number of tuples in D and Ci,D ,respectively. Let node N represent or hold the tuples of partition D. the attribute with the highest information gain is choose as the splitting attribute for node N. this attribute minimizes the information.

The expected information needed to classify a tuple in D is given by:

$$\mathrm{Info(D)} = -\sum_{i=1}^{m} p_i * \log_2(p_i) \qquad （1）$$

Where Pi is the probability that an arbitrary tuple in D belongs to class Ci and is estimated by | Di, d | / | D |.

Suppose we were to partition the tuples in D on some attribute A having Vdistinct values. As observed from the training data. If A is discrete-valued, these values correspond directly to the V outcomes of a test on A. Attributed A can be used to split D into V partitions or subsets. These partitions would correspond to the branches grown from node N. This amount of the new requirement is measured by:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} * Info(Dj) \qquad （2）$$

The item |Dj|/|D| acts as the weight of the jth partition. InfoA(D) is the expected information required to classify a tuple from D based on the partitioning by A.

Information gain is defined as the difference between the information requirement and the new requirement. That is:

$$gain（A）=Info(D)-Info_A(D) \qquad （3）$$

Choice highest information Gain(A), attribute A be taken as the splitting attribute *b) Gini index*

Uses the Gini index to be the measure of node impurity, the training data set T comes from m kind of example, the Gini target definition is:

$$\text{Gini(T)} = 1 - \sum_{i=1}^{m} p_i^2 \qquad (4)$$

We shall choice the smallest Gini property i as the splitting property of the node N.

c) Associated index based on the Goodman-kruskal

This is one kind method about the splitting measure carries on the digital measure in the finite division set when establishment decision tree has defined. This measure method has produced a coefficient GK source in Goodman-kruskal associated index [9], it had indicated that this kind of measure can succeed apply to decision tree's production.

Make $\pi = \{B1, \cdots, Bk\}$ and $\sigma = \{C1, \cdots, Cl\}$ are the two divisions of set S. The Goodman – Kruskal's coefficient of π And σ is:

$$GK(\pi, \sigma) = 1 -1 | S | \Sigma \; k \; i = 1 \; \max 1 \leqslant j \leqslant i \, | Bi \leqslant Cj | \qquad (5)$$

Structure measures dGK in a limit division set. Suppose the data set T = (T, H, rho) is uses in establishing decision tree J, assume v is the node that J will make divide, ρv will be the example set opposite with v, record on goal division in the data set ρv as $\theta\rho v$. Property Ai belonged to the property set H decide a relational ρ on division πA.

$$dGK(\pi A R, \theta R) = GK(\pi Ai \rho v, \theta \rho v) + GK(\theta \rho v, \pi Ai \rho v) \qquad (6)$$

Based on the dGK minimum value, cannot produce the precision good decision tree, but can use in choosing the splitting properties successfully, it establishes the decision tree is smaller and has the comparable precision.

2) TTVA apply to the crux issues mining

Let L denote the crux issues mining set's labeled sample set with size | L | and U denote the unlabeled sample set with size |U |. Three decision trees classifiers are respectively training in the use of sample set L after the data bootstrap sampling. After this training, each retraining behind, for any classifier, an unlabeled example can be labeled for it as long as the other two classifiers agree on the labeling of this example, while the confidence of the labeling of the classifiers are not needed to be explicitly measured. One of the three classifiers as the training classifier when the other two as complementary classifiers, the two complementary classifiers classify the samples set U, the same views with the sample set and the corresponding marking L, then to use |LUL' | training to the first category. The next round optimization process of the L does not act as labeled data but to be returned to U. Zhou [7] proved it to be overcome noise impact through this means.

Three decision trees voting classification algorithm based on Tri-training applied to the crux issues mining is as follows:

TTVA (L, U, Learn)

Input: L: Original labeled communications data set

U: Unlabeled communications data set

Learn: Learning algorithm (classifiers: H1,H2,H3)

Output: final result of rebelled S.

a) Initialization

(a) L0←L, L02←L, L03←L, S←L;

(b) M0←train (H1, L01);

(c) M02←train (H2, L02);

(d) M 03←train (H3, L03).

b) Circle

(a) use of sample set L of S after the data bootstrap sampling, produce three training set Li1 , Li2 , Li3 ;

(b) Ui←add unlabeled data from U ;

(c) rebelling from Mi1, Mi2, Mi3, according to splitting property;

(d) Politic choice labeled sample set {P1}, {P2} and {P3};

(e) create new training set of Hi: Li + 11←Li1 + {P1} ,Li + 12←Li2, {P2}, Li + 13←Li3 + {P3};

(f) Mi + 11←train (H1, Li + 11) , Mi + 12←train (H2, L i + 12) , Mi + 13←train (H3, Li + 13); S←Li + 11 + Li + 12 +Li + 13;

(g) Joint classifiers {H1, H2, H3} to reclassification the new rebelled data of S.

c) *Repeat the processes until U is empty.*

4. Experimental result and discussion

In this paper, the author’s software development company’s the year 06～08 data sets of communicated information are used in the experiments, and take the crux issues as Data Mining Research Object. They are two types property(Y/N). Since the end of the algorithm will be applied to the development and the teem members, in order to make more accurate and identify of the algorithms, randomly selected 4280 data records from the year 06～08 collected data set, the data is divided into two parts, 2020 records of training set and 2260 records of test set to modeling. In the experiments, the decision trees use the classics algorithm and respectively choice information gains, Gini index and associated index based on the Goodman-kruskal as the distinction eigenfunction of the three classifiers. The data is divided into three parts randomization, each part has the similar distributed condition to the whole sample proportion positive and negative, divide into two parts based on the different proportion in the experiment, one part data as labeled data set L, the other part data as unlabeled data set U. TTVA choice the TTVA error rate, the initial error rate, the recognition rate and accuracy rate of the classification algorithm regarded as estimate guideline of the algorithm performance.

TTVA analyze training data and test data to assess the accuracy of classification. And applied different number of unlabeled data in experiments to select the decision trees distinct initial error rate and the final TTVA error rate as a forecast of its performance of the algorithm in Table I to Table III.

TABLE I. 20% UNLABELED RATE

Error rate / Data set	TTVA (%)	Decision tree1		Decision tree2		Decision tree3	
		Initial (%l)	*Improve (%l)*	*Initial (%l)*	*Improve (%l)*	*Initial (%l)*	*Improve (%l)*
Section1	6.86	9.12	24.78	11.87	42.21	10.18	32.61
Section2	6.21	10.32	39.83	13.11	52.63	11.67	46.79
Section3	6.00	10.94	45.16	10.01	40.10	9.32	35.62

TABLE II. 50% UNLABELED RATE

Error rate / Data set	TTVA (%)	Decision tree1		Decision tree2		Decision tree3	
		Initial (%l)	*Improve (%l)*	*Initial (%l)*	*Improve (%l)*	*Initial (%l)*	*Improve (%l)*
Section1	7.10	13.43	52.79	12.59	43.61	12.49	43.15
Section2	6.34	12.02	47.25	13.31	52.37	13.32	52.40
Section3	7.82	13.48	41.99	15.76	50.38	13.17	40.62

TABLE III. 80% UNLABELED RATE

Error rate / Data set	TTVA (%)	Decision tree1		Decision tree2		Decision tree3	
		Initial (%l)	*Improve (%l)*	*Initial (%l)*	*Improve (%l)*	*Initial (%l)*	*Improve (%l)*
Section1	7.12	15.13	52.94	14.84	52.02	13.27	46.35
Section2	7.83	17.34	54.84	15.08	48.08	15.73	50.22
Section3	8.02	15.29	47.54	16.30	50.80	13.91	42.34

Tables I~III show the aspects of the classification error rate. With tri-training and different unlabeled rates of the data sets, the classification accuracy rate of TTVA has been greatly enhanced. And with the unlabeled rate increased the initial error is noticeably increased, but after TTVA the decision trees error rate marked a significant improvement, and also with the different three unlabeled rates the TTVA error rates are very close (all below 10%). This illuminates that TTVA can enhance the difference of the classifiers by use distinct eigenfunction. Then enhance the algorithm classification accuracy after integration and the output is very convenient. Only use a small number of labeled data and a large number of unlabeled data to correct the classifiers again and again. This algorithm overcomes the shortcomings of the single decision tree algorithm.

Apply TTVA to crux issue mining of the information communication. Use identification rate and accuracy rate to assess the classification algorithm. Identication rate and accuracy rate formula as follow:

$$Identification(R) = \frac{Ncorrect}{|D|} \tag{7}$$

$$Accuracy(R) = \frac{Ncorrect}{N\,\mathrm{cov}ers} \tag{8}$$

Ncovers denotes the forecast total correct, Ncorrect denotes the forecast real correct number, |D| denotes the actual correct total.

The integration of the classification algorithm shows in table IV.

TABLE IV. FORECAST APPLICATION RESULT OF YEARS 06~0 8 DATA WITH TTVA

Forecast / Real	not crux issues	crux issues	total
not crux issues	1510	48	1558
crux issues	28	674	702
Total	1538	722	2260

According to Table IV can obtain the forecast identify rate is 674/702=96.01%, the forecast accuracy rate is 674/722=93.35%.Looked from years 06~08 data test results show that its identification rate and accuracy rate achieved above 90%, these make out that the algorithm TTVA has met the mining requirements in view of years 06~08 data of communicated information. The example confirmed the identity and the accuracy of the information communication mode pattern to rely on the algorithm TTVA; the experimental results demonstrated the algorithm TTVA was suitable for the crux issues mining.

5. Conclusion

In this paper, a framework for intelligent information communication mode is designed, using Tri-training algorithm based on semi-supervised learning to improve the performance of mining. And design the three decision trees voting classification algorithm based on tri-training (TTVA), focusing on the use of the mining algorithm to the crux issues mining. Experimental results show that apply the algorithm TTVA to crux issues mining is realistic and can mine out the crux issues to help managers to make decisions. At the same time, TTVA overcomes the shortcomings of single tree algorithm which are computational inefficiency and supervised learning. Application data mining methods and share repositories to intelligent information communication mode of software development will greatly improve the speed of access knowledge and greatly reduce the rate of depletion of artificial information communication. Thereby improve software development efficiency and support information sharing and keep the active information communication. However, the study of intelligent information communication mode will be difficult, first of all, the diversity and complexity and timely changed of communicated data is not easy to collect; secondly, creating the data warehouse needs a certain amount of funds and personnel to support. At present, many software development companies have achieved questions collection system but intelligence is also far away. How to further information communication mode more intelligence is the direction that we continue to study.

References

[1] Vineeth Mekkat ,Ragavendra Natarajan , Performance characterization of data mining benchmarks.,TKDD, 2010.3

[2] Ted E. Senator,On the efficacy of data mining for security applications,International Conference on Knowledge Discovery and Data Mining,2009 .6

[3] Huang ming, Niu Wenying, Liang Xu, An improved decision tree classification algorithm based on ID3 and the application in score analysis[J], 2009 Chinese Control and Decision conference, 1876-1878.

[4] Carson Kai-Sang Leung , Efficient algorithms for mining constrained frequent patterns fromuncertain data, ACM 2009.6

[5] Damon Fenacci, Björn Franke , John Thomson, Workload characterization supporting the development of domain-specific compiler optimizations using decision trees for data mining ,SCOPES, 2010.5

[6] JiaweiHan,Michelin Kamber, Data Mining Concepts and Techniques,Second Edition.

[7] Zhi-Hua Zhou,Ming li,Tri-training: Exploiting Unlabeled Data Using Three Classifiers,IEEE Trans on knowledge and Data Engineering, 2005.17(11).

[8] Shirish Tatikonda ,Srinivasan Parthasarathy ,Mining tree-structured data on multicore systems.,Proceedings of the VLDB Endowment,2009.8

[9] TheWinPcapTeam,WimPcnp,Documentation[EB/OL],http://www.wimpcap.org/docs/docs31/html/main.html,2005.12

[10] Huang Yuanhang,Liu Hongwei,Software development from information communicatiom mode,Computer Applications and Software,.2007-02. (in Chinese).

[11] Liujinxi, the evolvements and the Translations of information communication mode. journal of modern information,2010.5.(in Chinese).

[12] Zhudeli,SQL Server 2005 and data mining and bussiniss intelligentize entirety resolve ,publishing house of electronics industry.2007.10 w(in Chinese).

[13] chen wenwei, Data warehouse and data mining education. tsinghua university press,2006. (in Chinese).