Data Mining Approach in Software Analysis

S.Suyambu Kesavan
Research Scholar
Sivanthi Aditanar College
PillayarPuram, Nagercoil

Dr.K.Alagarsamy,
Associate Professor
Dept of MCA, Computer Center,
Madurai Kamaraj University, Madurai

Abstract:

Data mining and knowledge discovery have proved to be valuable tools in various domains such as production, health care and management. Data mining also has potential to address some highly challenging areas of software engineering such as adaptability and security. In software engineering process analyst play an important role for gathering information from the statement of user and obtaining the information from many resource. Data mining gives the potential algorithms and resource for collecting the information. In this paper we are merging the concept of data mining algorithms into software engineering techniques to collect the information and produce the better analyst decision.

Keywords: *Software engineering, Data Mining*, *Clustering algorithm.*

Introduction:

Data mining:

Data Mining, also popularly termed as Knowledge Discovery in Databases, refers to the nontrivial extraction of implicit, previously unprocessed and potentially useful information from data in databases. While data mining and knowledge discovery in databases are frequently treated as synonyms, data mining is important part of the knowledge discovery process. The Knowledge prediction in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- ***Data cleaning****:*

  It is a phase in which noise data and irrelevant data are removed from the collection.

- ***Data integration****:*

  At this stage, multiple data sources, often heterogeneous, may be combined in a common source.

- ***Data selection****:*

  The data relevant to the analysis is decided on and retrieved from the data collection.

- ***Data transformation***:

  It is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

- ***Data mining****:*

  It is the critical step in which clever techniques are applied to extract patterns potentially useful.

- ***Pattern evaluation:***

  In this step, strictly interesting patterns representing knowledge are identified based on given algorithmic procedure.

- ***Knowledge representation****:*

  This is the final phase in which the discovered knowledge is visually represented to the user. This step uses visualization techniques to help users understand and interpret the data mining results.

It is usual to combine some of these steps together. For instance, data cleaning and data integration can be merged together as a pre-processing phase to generate a data warehouse. Data selection and data transformation can also be merged where the consolidation of the data is the result of the selection or as for the case of data warehouses; the selection is done on modified data.

***Software Engineering*:**

Software Engineering is a technique dedicated to designing, implementing, and modifying software so that it is become of high quality, affordable, maintainable and fast to build. It is a step by step approach to the analysis, design, assessment, implementation, test, maintenance and reengineering of software development that is the application of engineering to software.

There are three main activities to be performed before to the start of software like planning, creating the stakeholder requirements, and defining and deploying the development environment.

Once these activities are completed, we are ready to initiate the project. The project in run as a series of incremental development efforts, each expanding and elaborating on the efforts that came before. Some of the important points

.

- ***Project Initiation***
  Setting up the team along with development environment, Understanding what the customer needs, planning the project
- ***Perspire planning***
  Parallel activities may be done by different people; Parallel activities may be done in any order with respect to each other.
- ***Creating the Schedule***
  Steps involved in this task are identifying the desired functionality, identifying the key risks
- ***Creating the team work***
  There exists a strong correlation between the team structure and the model organization. Teams are formed because they make coherent sense and model is organized to allow the teams to work together effectively.
- ***Planning for reuse***
  Identifying reuse needs and goals, Identifying opportunities for reuse, Estimates the cost of constructing reusable assets, determining which reusable assets to construct
- ***Planning for risk reduction***
  steps involved are Identify the key project problems, Quantifying problem severity, Determining the likelihood of these key project problems, Computing the project risks,
- ***Specifying logical architecture***
  This is also known as project structure or model organization involving specification of model organization patterns and checklist for logical architecture.
- ***Performing the initial safety and reliability analysis***
  Steps involved are Identify the hazards, quantify the problems in terms of likelihood and severity, Compute the risks and perform an initial safety analysis.

These are the some of the key points where software analyst play a vital role for collecting the information. Information gathering is a huge process and tedious process. Information source may any form like document, ppt, pdf, etc. If it is manual search it may lead to some problem and accuracy problem due to that we are moving some other automation technique.

Data mining for SE:

Mining software engineering data has emerging as a research direction over the past years. This research direction has already achieved substantial success in both research and practice. In this paper, we declare Software Intelligence as the future of mining software engineering data, within modern software engineering research, practice and education. The vision of Software Intelligence (SI)has yet to become a reality. Nevertheless, recent advances in the Mining Software Repositories field show great promise and provide strong support for realizing SI in the near future, as software engineering research aims to ensure its relevance and impact on modern software practice. This position paper summarizes state of practice and research of SI, and lays out future research directions of mining software engineering data tenable SI.

Text mining is a new and exciting research area that attempts to solve the information overload problem. It uses many techniques from data mining, but since it deals with unstructured data, a major part of the text mining process deals with the crucial stage of preprocessing the document collections. The process also involves the storage of the intermediate representations, techniques to analyze these intermediate representations. A typical text mining system begins with collections of raw documents, without any labels or tags. Documents are then automatically tagged by categories, terms or relationships extracted directly from the documents. Next, extracted categories, Entities and relationships are used to support a range of data mining operations on the documents. In this paper we are going to implement the innovative approach of data mining in software engineering for helping he software analyst.

Proposed method:

In our research we are going to incorporate the probabilistic clustering method for our analyst usage. This is the combination approach of software engineering analysis phase with data mining clustering approach. Here we will discuss about the characteristics probability clustering approach, data is considered to be a sample independently taken from the user objectives of several probability distributions. The main idea is that data points are generated by first randomly taken a point x from a corresponding distribution. The area around the mean of each distribution constitutes a natural cluster. So we associate the cluster with the corresponding distributions parameters such as mean, variance etc. Each data point carries not only its attributes but also a cluster ID .Each point x is assumed to belong to one and only one cluster and we can estimate the probabilities of the assignment.

Customer objective serves as an objective function, which gives rise to the Expectation Maximization (E-M) method is a two-step iterative optimization. Step (E) estimates probabilities, which is equivalent to a soft reassignment. Step (M) finds an approximation to the customer objective given current soft assignments. This boils down to finding mixture model parameters that maximize estimation result. The process continues until estimation convergence is achieved.

We can also use some other tricks to facilitate finding better local optimum suggested acceleration of EM method based on a special data index, decision tree, KD-tree, etc. Data is split at each node into two descendants by dividing the widest attribute at the center of its range. Every node stores sufficient statistics. Approximate computing over a pruned tree accelerates EM iterations.

Probabilistic clustering has some important features:

- It can be altered to handle recodes of complex objective of the user.
- It can be stopped and resumed with consecutive batches of data, since clusters have representation totally different from sets of points
- At any stage of iterative process the intermediate mixture model can be used to assign cases (on-line property)
- It results in easily interpretable cluster system

Because the mixture model has clear probabilistic foundation, the determination of the most suitable number of clusters k becomes a more tractable task. From a data mining perspective, excessive parameter set causes over fitting, while from a probabilistic perspective, number of parameters can be addressed within the Bayesian framework.

Result Discussion:

We have taken set of sample user objectives of the software engineering for analysis purpose. We tested performance effectiveness with the human analyst with our probabilistic clustering method. Below diagram shows the effectiveness of the probabilistic clustering method.
The result value is given out of ten.

	Analyst	Proposed Method
Missed Requirements	4	2
Accuracy	6	9
Remembrance	4	8

Trace Back	3	9
Final Result	6	9

Table1

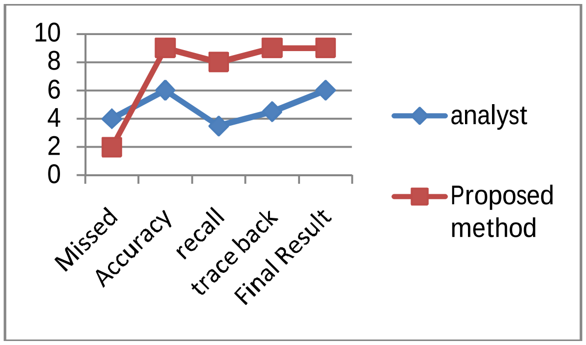

Fig:1

Conclusion:

In past decade, lot of innovative approach method has been implemented in software engineering and data mining, which yields the highly valuable solutions for software industry. In our research we carried out the data mining with software engineering approach for analyst purpose. Compare with human analysis it will give more accuracy and also it reduces the human effort and counting. It's our belief, our work will highly helpful to the software engineering society.

Future Work:

In our research we carried out probabilistic clustering method to form a clustering analysis of user objective. There are many clustering algorithm present in data mining domain. In future we can use some other algorithm for the implementation and deriving the conclusion from comparative analysis. We can also predict which one is the best algorithm for software engineering.

Reference:

[1]A. E. Hassan, A. Mockus, R. C. Holt, and P. M. Johnson.Guest editor's introduction: Special issue on mining softwarerepositories. *IEEE Trans. Softw Eng.*, 31(6):426–428, 2005.

[2] Josh Eno, Craig W Thompson," Generating Synthetic Data to Match Data Mining Patterns", IEEE Internet Computing May/June 2008 pp.78 – 82.
[3]. O.Maqbool, A Karim, H.A.Babri, Misarwar, "Reverse Engineering using Association Rules", IEEE INMIC 2004, pp. 389 -395.

[4]. Gang Kou Yipeng, "A Standard for Data Mining based Software Debugging", IEEE 4 The International Conference on NetworkedComputing and advanced Information Management, pp. 149 – 152.

[5]. Ray-Yaung Chang, Andy Podgurski, Jiong Yang, "Discovering Neglected Condition in Software by Mining Dependency Graphs",", IEEETransactions on Software Engineering, Vol. 34, No. 5, September/October 2008, pp. 579-596.

[6] Jensen, C. and Scacchi W. Simulating an Automated Approach to Discovery and Modeling ofOpen Source Software Development Processes. InProceedings of ProSim'03 Workshop on SoftwareProcess Simulation and Modeling, Portland, OR May2003.

[7] D. Harel and S. Maoz. Assert and negate revisited: Modal semantics for umlsequence diagrams. *Software and System Modeling*, 2007.
[8] D. Lo, S.-C. Khoo, and C. Liu. Efficient mining of iterative patterns forsoftware specification discovery.In *KDD*, 2007.