DATA MINING TECHNIQUES FOR EFFICIENT
SOFTWARE ENGINEERING

A. JANUPRIYA¹, R. HEMAMALINI², M. INBAVALLI³

¹,² MCA, Er. Perumal Manimekalai College of Engineering, Hosur
³ Associate Professor, MCA, Er. Perumal Manimekalai College of Engineering, Hosur

Abstract—Software Engineering is the computing field concerned with designing, developing, implementing, maintaining, and modifying software. Software Engineering data consists of sequences, graphs, and text. Meaningful information can be extracted from the complex data using well-established data mining techniques such as association, classification, clustering etc. The potential of data mining technique on this large valuable data is to manage the software projects efficiently and to produce high-quality software systems that are delivered on time and within budget. Data mining for secure software engineering improve software productivity and quality. Recently software engineers are increasingly applying data mining algorithms for various software engineering tasks. This paper focuses on data mining techniques for software engineering process[1].

Keywords—Software Engineering, Software Engineering goals and tasks, Data Mining, Mining Software Engineering data.

I. INTRODUCTION

Software development is to develop quality software. The quality software is developed using vast amount of software engineering data. The software engineering data can be used to gain empirical based understanding of software development. It can be used to predict, plan and understand various feature of the project. Software engineering data can support future growth and project management activities[1]. Software engineering data (such as code bases, execution traces, historical code changes, mailing lists, and bug databases) contains a wealth of information about a software project’s status, progress, and evolution. But this data is scattered. Meaningful information can be extracted from this data using some data mining techniques such as classification, association, pattern matching, clustering etc., and using certain data mining tools[3]. DM is used to locate required information and facts which are stored in different database. DM is the process of analysing data from different perspective and summarizing it into useful information. Information that can be used to increase revenue, cut costs, or both. DM software is one of a number of analytical tools for analysing data. DM tools are useful in predicting the future trends and behaviour which are helpful for engineers to take proactive knowledge driven decisions[4].

II. LITERATURE REVIEW

In [1] Applications of DM Techniques in Software Engineering is discussed as some of the data mining techniques which can be applied to different types of software engineering data in order to solve the challenges posed by software engineering tasks such as development, management, debugging, and maintenance. Also data mining is as good as results it produces so quality and quantity of available data. In [2] Application of DM Techniques for Improving Software Engineering is discussed as the computing field concern with designing, developing, implementing, maintaining, and modifying software. In this paper, we study how data mining techniques can be applied in solving Software Engineering problems. In [3] Application of DM Techniques on Software Engineering Data for Software Quality is discussed as the processes of Software engineering are complex and produces large number and variety of artefacts. The potential of data mining techniques on this large valuable data is to better manage the software projects and to produce high-quality software systems that are delivered on time and within budget.
III. GOAL OF SOFTWARE ENGINEERING

The following are the goals of software engineering. Data mining is quite useful in achieving these goals efficiently and quickly as well.

(a) To find and fix bug: Software bug estimation is a very essential activity for effective and proper software project planning. All data related with software bug is kept in software bug repositories. A software bug repository contains interesting information related to the development of a project. Data mining techniques can be applied on these repositories to uncover useful and interesting patterns. A prediction data mining technique is used to predict the software bug estimation from a software bug repository.

(b) Documentation: Software document data is high important but it is complex in nature processed by data mining techniques. Source code, system administration and application documentation consists large buffer of documents and free text for mining and software analysis. External and internal documentation also plays the important role for data sources. The types of documents (html, portable document format, text etc.) available in large variety and another important source are the multimedia data (audio, video figures).

(c) Software configuration management data: Software configuration management a system (SCMs) includes documents, software code, status accounting, design models, defect tracking and also include revision data. In SCMs the large amount of data is available and most valuable data is kept from different sources.

(d) Source code: Source code is an important source for data mining in software engineering. The various applications of data mining in software engineering is program comprehension, maintenance and software components analysis. Initially available source code is always a caveat and parses source code language available and it can be seen as structured form. Applying data mining techniques in source code includes predicting change propagation, change history, predicting defect densities in source code files.

(e) Mailing lists: Large software systems especially open source software bridging developers and users. Mailing lists contain hard data contain a lot of free message, text and author graphs. Data mining applications are not limited to text analysis, linguistic analysis and text clustering of subjects.

(f) Cost estimation: Cost estimation is approximate judgments of the cost for a project and it is the one of the main problem in software engineering there are too many variables involved to calculation for the cost estimate (technical, human, political and environmental) and measure in terms of efforts and metric used is person months or year. Accurate cost estimation is very important for every kind of project and estimated by COCOMO model [6].

IV. DATA MINING TECHNIQUES

4.1 Association rule

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Association Rule mining techniques is applied to the records in order to discover the patterns that are likely to cause high severity defects. The discovered patterns are then helpful to predict the subsequent actions that may result in high severity defects. The concept of association in mining software engineering data is based on set of strong rules. The computational cost of association rules mining in software development can be reduced by reducing the number of passes over the database, sampling the software database, adding extra constraints on the structure of patterns, through parallelization.
4.2 Classification

Classification is one of the main tasks in data mining for assigning a data item to a predefined set of classes. Classification can be described as a function that maps (classifies) a data item into one of the several predefined classes. Here the goal is to induce a model that can be used to classify future data items with unknown classification into unique classes. In software development process the performance of classifier depends upon the type and class of data. There are different forms of data available in software engineering. It is imported to work with relevant data items and classify them into sub classes and keep on adding new data items into pre-existing classes. We implement different classification algorithms in software engineering to solve various problems in different phases. Classification can be used to identify the types of bugs and thus helps in building bug detector.

4.3 Clustering

Cluster analysis is a group of multivariate techniques whose primary purpose is to group entities based on their attributes. Similar objects are placed in the same cluster according to predetermined selection criteria. The objective of any clustering algorithm is to sort entities into groups, so that the variation between clusters is maximized relative to variation within clusters. The set of entities to cluster needs to be identified, before applying clustering to a software system. Clustering algorithms used in software engineering are: graph-theoretical algorithms, construction algorithms, optimization algorithms, hierarchical algorithms.

4.4 Text mining

Approximately 80% of information is stored in computers is in form text. Example of software engineering text data includes project and bug reports, e-mails and code comments. Code duplication is one of the biggest problems which complicates maintenance and evolution of software systems. Several drawbacks of all existing code duplication techniques can be overcome by using visual approach which is language-independent. Text data mining refers to the discovery of hidden information and potentially useful knowledge from a collection of texts which is done by automatic extraction and by analysing information [7].

Table 1. Data Mining techniques in Software Engineering [2].

<table>
<thead>
<tr>
<th>Sr. NO</th>
<th>DATA MINING METHODS AND TECHNIQUES</th>
<th>FIELD OF APPLICATION IN SOFTWARE ENGINEERING</th>
<th>TYPE OF DATA SET</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Association rule</td>
<td>Catalog design, prediction of severe defects, cross marketing.</td>
<td>Large variable data set, numerical, alphanumeric, visual and audio data.</td>
</tr>
<tr>
<td>2.</td>
<td>Clustering Technique</td>
<td>Development of cost effective tools, discovery and localization of program failures.</td>
<td>Statistical data, discrete and comparative numerical data.</td>
</tr>
<tr>
<td>4.</td>
<td>Text mining</td>
<td>Detection of code duplication, bug duplication reports.</td>
<td>Project reports, bug reports, codes, emails, text data.</td>
</tr>
</tbody>
</table>

V. DATA MINING TECHNOLOGY IN SOFTWARE DEVELOPMENT PROCESS

In recent years, data mining technology is widely used in software engineering, and applying
data mining in software engineering can improve the maintenance efficiency of software system, which also to a certain extent, increases system stability.

5.1 Data mining in programming

Programming, as an important content of developing software, writes the code. Developers need to fully understand the structure and function of programming code, according to their own understanding, select valuable information in the database. Usually programming needs information of following parts:

- Software developers find needed code structure, similar function, and patterns which can be reused, such as data structures, object, method and so on in the existing code library;
- Developers can find static rules for reusing some patterns in the database, for example: class method, inheritance relationships, and so on;
- Developers have a thorough analysis understanding of reuse pattern rules, such as: API call order [9].

5.2 Data mining of developing open source software

Open source software refers to source code development software, which is a free provided service for customer. For it is free, the management and control of open source software is more difficult. At this point, data mining technology can improve the quality of open source software. For example, data mining system designed at the University of Oxford, users can real-time track and manage system, so to a certain extent, it improves use efficiency of open source software.

5.3 Data mining used in program code

Program code refers to clone code, which is a reusable code using copy and paste operations. Data mining used in cloning code is tested early. Cloning code testing mainly adopts the following forms: text comparison method, method based on measurement, latent semantic indexing, etc. But within clone code, using data mining is not enough mature, mainly because data mining must take consideration of semantic mining. At the same time, digging to the crosscutting concerns, crosscutting concerns use more mining methods, for example: in the process of code text analysis, according to different characteristics, divide into analysis based on text and type, based on clustering analysis, formal concept analysis, etc.

5.4 Data mining technology using for software fault detection

Data mining technology can mine to program code and interactive modes according to the procedure of enforce discipline mining, thus accurate position and detect software failure. Automatic mining is a more mature API rule mining method. Traditional positioning software failure uses program slicing, which is more complex, prone to fault and location is not allowed. As traditional positioning method is gradually improved, program spectrum abstract describe trajectories is more successfully used, and compare the running of software with fault running state, according to the difference of the two, judge fault sources, and these new data mining technologies can effectively improve the accuracy and efficiency of software fault detection.

5.5 Data mining technology in software management

In software project management, data mining technology is used mainly in organizational relationships and version control information. Software project management is a complex project, and the key lies in reasonable personnel organization relationship mining are coordination and allocation of human resources, such as: a project may have hundreds of thousands of people involved in this process, which involves various personnel interactions, which are implemented via E-mail, documents and other interactions. Version control can record the change of the whole file content in
detail, making it easy for users to view the revised version, and apply data mining technology to the late version control information can reduce system maintenance cost. Data mining can provide warning role for daily maintenance software system in time. Some mining data can clearly looking for errors existing in system repair process, according to mistakes records, facilitate software designers to avoid the common mistakes in time, and promote the restoration and management level of software project [5].

VI. SOFTWARE ENGINEERING TASK THAT BENEFIT FROM DATA MINING

Role of data mining in improving effectiveness of software engineering tasks.

6.1 Development tasks
Software development is a creative process as no two programs are the same. It is difficult to accumulate enough relevant data in the initial programming phase of a software project which provide insights that is helpful in guide development. With use of dynamic analysis and mining of revision histories, bugs can be fixed with constant check-ins. The errors found using this approach in development phase are mostly previously unknown.

6.2 Management tasks
Managers can utilize the historical data and software artefacts to improve the management task. It becomes even important for manager dealing with extremely large projects as problems such as bug prediction and resource allocation arise and data mining provides many innovative solutions. Use of software tools improves the quality of software but for large projects and organizations it becomes expensive and hard to manage and maintain. Data mining helps in cost based analysis and selection of proper tool depending upon the tool usage statistics with estimates of developer effort.

6.3 Research tasks
The goal of data mining from point of view of engineering researcher is to gain understanding about a number of projects which is helpful in characterizing patterns in software development. Researchers generally analyse data from open-source projects, but mining data from organizations like Sourceforge.net is fraught with drawbacks such as dirty data and defunct projects. Software evolution is favourite and latest topic for software data miners. A better way to understand a program development history is by making use of partitioning and clustering of version data[8].

![Figure 1 Overview of Data Mining in Software Engineering](image-url)

Figure 1 Overview of Data Mining in Software Engineering[10].
VII. CONCLUSION

We have discussed the application of data mining for solving of a number of Software Engineering problems. A number of problems encountered in the SE field, such as bug occurrence, high cost of software maintenance; unclear requirements and so on can reduce software quality and productivity. We outlined the data mining techniques that can be applied to various types of SE data in order to solve the challenges posed by Software Engineering tasks such as programming, bug detection, debugging, and maintenance. Our studies show that data mining techniques have proven to be effective for improving Software Engineering by increasing software reliability and quality.

As to future direction, new mining techniques or algorithms are needed in order to solve unclear software requirements in mining SE data. Another challenge is the ability to mine combination types of SE data (e.g., text and graphs together) for more informative patterns. Further research needs to be done to explore the potential of mining combination data to solve numerous real-world problems in Software Engineering.

REFERENCES