A STUDY OF APPLICATION OF NEURAL NETWORK TECHNIQUE ON SOFTWARE REPOSITORIES

Ana Maria Bautista, Tomas San Feliu

Abstract: Context: The use of software repositories is recent in software engineering. Different techniques have been applied to software engineering problems. We want to know the degree of application of neural networks as data mining technique. The results have allowed the preliminary state of the art of the application of neural network techniques on software repositories.

Keywords: neural networks, repository, data mining.

ACM Classification Keywords: I.2.6 Artificial Intelligence –Connectionism and neural nets, H.2.7 Database Administration – Data warehouse and repository.

Introduction

The main goal of a software repository is to maintain the project data for later use. The data in mature engineering are used to check and compare different techniques.

In software engineering there are now a wide variety of software repositories. Software engineering as an emerging discipline is beginning to use the software repositories to compare different techniques.

The systematic review can be considered as a secondary study that reviews articles related to a specific search topic. As a secondary study provides an overview of an area of research to evaluate existing evidence [Kitchenham, 2009] and can provide identifying gaps in primary studies, where they may require new or enhanced studies.

These studies require a rigorous search and inclusion criteria and exclusion that are clearly defined in the research protocol and are presented in the report results.

The purpose of this study is to determine which use is made of the repositories and techniques have been tested and compared with them.

A software engineering (SE) data repository is defined as a set of well-defined, useful, and pertinent real-world data related to software projects, called datasets, which include quantitative and descriptive information about resources, processes, techniques, management, etc. Such data are being collected for various purposes by recognized organizations, as well as by individual software organizations and researchers. In most scientific and engineering disciplines, these data are useful for conducting benchmarking, experimental, and empirical studies. While highly varied and widely available in mature disciplines, data repositories are much less frequently found in emerging disciplines, including software engineering, as illustrated by the Guide to the Software Engineering Body of Knowledge [SWEBOK, 2004].

Mining software repositories (MSR) has become a fundamental area of research for the Software Engineering community, and of vital importance in the case of empirical studies. Software repositories contain a large amount of valuable information that includes source control systems storing all the history of the source code, defect tracking systems that host defects, enhancements and other issues, and other communication means such as mailing lists or forums.
To extract information from the Software Repositories different techniques are used. Mohanty et al. classify intelligent techniques in the following [Mohanty, 2010]:

1. Different neural network (NN) architecture including multilayer perception (MLP) and cascade correlation NN;
2. Fuzzy logic;
3. Genetic algorithm (GA);
4. Decision tree;
5. Case-based reasoning (CBR);

The other techniques:
1. Analogy based;
2. Support vector machine;
3. Self organizing maps (SOM).

Specifically, this work will focus on studying the application of neural networks in existing repositories.

Neural networks are used broadly in the studies we have selected. Therefore consider your extension, we will set the yields obtained in the literature and discuss the tendency in recent years.

This paper will be organized as follow. Research methodology section will describe the systematic review. Next section title Data Collection, reports the most relevant information gathered. Results section will report the review results analyzing collected date ordered by research questions. Discussion section summarizes the main findings. Study limitation section will discuss the assumptions and considerations of the study. Finally, conclusion and future work will outline the main conclusions obtained in future research works

Research methodology

This section provides an overview of the steps involved in the process systematic review, including the formulation of research question, the search strategy, the inclusion and exclusion criteria, and finally the data collection process.

Systematic mapping studies are a type of systematic literature review that aims to collect and classify research papers related to a specific topic [Petrie, 2006; Kitchenham, 2007; Petersen, 2008].

This section provides an overview of the steps involved in the process of mapping review following Petersen et al. [Petersen, 2008] including the formulation of the research questions, the search strategy for primary studies, the inclusion and exclusion criteria, and the data collection process.

• Research questions

The main goal addressed by this study is to analyze the use of repositories and techniques by the research community and to consider its weakness to undertake the appropriate scientific research.

In this study the following research questions were considered:

✓ **Research question number 1 (RQ1):** Which and how many journals and conferences include techniques for mining software repositories research papers? To identify what are the main literature sources where the software repository analysis are published.
✓ **Research question number 2 (RQ2):** How comprehensive is the use of neural networks in the analysis of software repositories? Is necessary to know the use of neural networks for the analysis of repositories and is required to characterize the types of networks are used.
Research question number 3 (RQ3): In which years have conducted studies with neural networks? How have been evolving the use of neural networks over time. The aim is to establish whether there is a trend over the last years.

• Search for primary studies strategy

The following search engine Google Scholar was used to make a general search for relevant papers in journals and conference proceedings. This search engine was selected because it is major search engine and it has a good usability.

Search is based in the two most used open software repositories: ISBSG and PROMISE. Search terms Data Mining ISBSG Repository and “Data Mining” “PROMISE Repository” were used. Only have been considered papers published from year 2010. Search was completed in January 2014. There were 304 papers, 94 corresponding to the first search term and 210 to the second.

Some studies use both repositories and therefore they are being selected for both search terms. Also it was noted that some papers have been published in different journal and conference proceedings or in different years. Duplicate references have been eliminated, overall 46 papers.

Finally, 258 references remained. The overall primary study selection is summarized in Figure 1.

Figure 1. Search process for the selection of studies

• Inclusion and exclusion criteria

Inclusion and exclusion criteria are required to evaluate each primary study. In order to improve its reliability, only one author performed the screening process.

Inclusion and exclusion criteria are required to assess each potential primary study. In order to improve its reliability, the filtering process based on inclusion and exclusion criteria. All conflicts were resolved via discussion.

The following list describes the filtering process:
F1: The first filter (F1) was used to identify papers that didn’t speak about data mining techniques and for this reason they didn’t answer our search questions, 82 papers were filtered;

F2: The second filter (F2) was used to identify reviews that didn’t anything new to the rest of papers, 5 papers were filtered.

F3: The third filter (F3) was used to locate the papers that were not written in English, 2 papers were filtered.

169 papers satisfied the logical condition (F1 AND F2 AND F3). Filtering process is summarized in Figure 2.

Data Collection

After the filtering process, the most relevant information was obtained from remained studies.

After the filtering process, the most relevant information was obtained from each of the 169 remained studies. This includes both general information and data addressing the five research questions. After reading each paper, the data was extracted and stored in a spreadsheet.

Papers classified by sources are shown in Table 1. Table 1 shows the evolution along the filtering process.

<table>
<thead>
<tr>
<th>Source</th>
<th>Initial dataset</th>
<th>Duplicates</th>
<th>Filtered</th>
<th>Final dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE</td>
<td>81</td>
<td>12</td>
<td>20</td>
<td>49</td>
</tr>
<tr>
<td>ACM</td>
<td>29</td>
<td>9</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>Springer</td>
<td>15</td>
<td>2</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Others</td>
<td>179</td>
<td>23</td>
<td>56</td>
<td>100</td>
</tr>
<tr>
<td>TOTAL</td>
<td>304</td>
<td>46</td>
<td>89</td>
<td>169</td>
</tr>
</tbody>
</table>

It was analyzed the filtering process from the point of view of publication year. This point of view is shown in Table 2.
Table 2. Papers by published year

<table>
<thead>
<tr>
<th></th>
<th>Initial dataset</th>
<th>Duplicates</th>
<th>Filtered</th>
<th>Final dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>60</td>
<td>5</td>
<td>24</td>
<td>31</td>
</tr>
<tr>
<td>2011</td>
<td>51</td>
<td>6</td>
<td>13</td>
<td>32</td>
</tr>
<tr>
<td>2012</td>
<td>70</td>
<td>9</td>
<td>24</td>
<td>37</td>
</tr>
<tr>
<td>2013</td>
<td>108</td>
<td>25</td>
<td>27</td>
<td>56</td>
</tr>
<tr>
<td>2014</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>TOTAL</td>
<td>304</td>
<td>46</td>
<td>89</td>
<td>169</td>
</tr>
</tbody>
</table>

Additionally, the different sources was analyzed the distribution of type of papers. The most important source is the journals. The conferences are the second source of information. The Table 3 and Table 4 present the distribution of studies across different journals and conferences.

Table 3. Papers by Journal

<table>
<thead>
<tr>
<th>Journal</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information and Software Technology</td>
<td>11</td>
</tr>
<tr>
<td>IEEE Transaction on Software Engineering</td>
<td>9</td>
</tr>
<tr>
<td>International Journal of Software Engineering and Knowledge Engineering</td>
<td>7</td>
</tr>
<tr>
<td>Empirical Software Engineering</td>
<td>6</td>
</tr>
<tr>
<td>Journal of Systems and Software</td>
<td>4</td>
</tr>
<tr>
<td>Information Sciences</td>
<td>4</td>
</tr>
<tr>
<td>Software Quality Journal</td>
<td>3</td>
</tr>
<tr>
<td>IET Software</td>
<td>2</td>
</tr>
<tr>
<td>IEEE Transaction on Reliability</td>
<td>2</td>
</tr>
<tr>
<td>International Journal of Software Engineering and Computing</td>
<td>2</td>
</tr>
<tr>
<td>International Journal of Reliability, Quality and Safety Engineering</td>
<td>2</td>
</tr>
<tr>
<td>Others</td>
<td>36</td>
</tr>
<tr>
<td>TOTAL</td>
<td>88</td>
</tr>
</tbody>
</table>

Table 4. Papers by Conference

<table>
<thead>
<tr>
<th>Conference</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>International Conference on Predictive Models in Software Engineering</td>
<td>8</td>
</tr>
<tr>
<td>International Conference on Software Process and Product Measurement</td>
<td>6</td>
</tr>
<tr>
<td>CSI International Conference on Software Engineering</td>
<td>3</td>
</tr>
<tr>
<td>International Joint Conference on Neural Networks</td>
<td>3</td>
</tr>
<tr>
<td>IEEE/ACM International Conference on Automated Software Engineering</td>
<td>2</td>
</tr>
<tr>
<td>International Conference on Computer, Information and Telecommunication Systems</td>
<td>2</td>
</tr>
</tbody>
</table>
One important question is related to the occurrence of neural networks in data mining software repositories. It was gathered information about the techniques used in selected studies. Sometimes one study contains several techniques. The distribution of techniques used is shown in Table 5.

### Table 5. Distribution of techniques

<table>
<thead>
<tr>
<th>Technique</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural Networks</td>
<td>14.29</td>
</tr>
<tr>
<td>Fuzzy Logic</td>
<td>2.77</td>
</tr>
<tr>
<td>Genetic algorithm</td>
<td>0.85</td>
</tr>
<tr>
<td>Decision trees</td>
<td>17.48</td>
</tr>
<tr>
<td>Case-base Reasoning</td>
<td>2.13</td>
</tr>
<tr>
<td>Hybrid intelligent systems</td>
<td>5.97</td>
</tr>
<tr>
<td>Analogy based</td>
<td>5.12</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>2.77</td>
</tr>
<tr>
<td>Self organizing maps</td>
<td>2.77</td>
</tr>
<tr>
<td>Statistical techniques</td>
<td>44.78</td>
</tr>
<tr>
<td>Others</td>
<td>1.07</td>
</tr>
</tbody>
</table>

### Results

The main results of systematic review are presented following the researched questions:

- **RQ1:** Which and how many journals and conferences include techniques for mining software repositories research papers?

  The publications of data mining techniques in software repositories have been captured from IEEE and ACM. The publications are concentrated in the domains of Software Engineering and Knowledge Engineering. In Table 4 is shown the ranking of journals. The three main journals are Information and Software Technology, IEEE Transaction on Software Engineering and International Journal of Software Engineering and Knowledge Engineering. The most important conference in this field of study are International Conference on Predictive Models in Software Engineering, International Conference on Software Process and Product Measurement, CSI International Conference on Software Engineering and International Joint Conference on Neural Networks. The tendency established by the journals is confirmed by the conference. It is denoted IJCNN is specialized in neural networks and listed more conference publications.

- **RQ2:** How comprehensive is the use of neural networks in the analysis of software repositories?
This question is required to identify the use of neural networks in the studio of repositories and it requires characterize the types of networks are used. In the 169 papers selected in our study all kinds of data mining techniques are used. In many cases hybrid techniques are used with the original techniques or algorithms that improve their performance complement. In all studies the performance of at least two techniques are compared and in many numbers much higher. Most frequently statistical techniques are used. Neural Networks are ranked in third position below of decision trees as shown in Table 5. The multilayer perceptron (MLP) and Radial Basis function Network (RBFN) are some of the popular Neural Network architecture.

- **RQ3: What years there have been conducted studies with neural networks?**

How have been evolving the use of neural networks over time? It intends to establish whether there is a trend over the last years. The figure 3 shows that there is a significant increase over 2013.

![Neural Networks Papers](image)

**Figure 3. Evolution of Neural Networks publications**

**Discussion**

This section summarizes the main findings of the systematic review. It also includes the limitations of the study and discusses the implications for researchers.

This study shows the extent to which and how software engineering researchers have used ISBSG and PROMISE until January of 2014. Thereby, the papers that have worked with these datasets have been identified and classified by answering a set of research questions. This systematic mapping review conveys a picture of application of neural networks on software repositories.

The search terms Data Mining ISBSG Repository and “Data Mining” “PROMISE Repository” were the input for Google Scholar. This resulted 94 y 210 results respectively. The search was completed in January 2014. After the elimination of duplicates and the filtering process, the most relevant information was obtained from each of the 169 remaining studies.

The first research has been a consolidation of journals and conferences where articles are published on this subject. It is noteworthy that the conference that has emerged in recent years the subject of analysis software repository is maintained.

A second research question shows that the neural networks maintain their presence as data mining technique. Furthermore, from the viewpoint of performance which is a consolidated sample technique. In fact neural networks are consolidated as reference technique in comparisons to characterize its performance.
Regarding the third question, we see that during the years 2010 to 2012 remained stable presence. In the years 2013-2014 a remarkable growth over the previous three years is provided.

**Study limitations**

It is important to consider that the results obtained form a systematic review could be affected by researchers conducting the review, by the selected search terms, and by the chosen time frame [Elberzhager, 2012].

The first limitation concerns the search strategy employed, to use ISBSG and PROMISE in search terms, results can be affected.

It is important to indicate that some more recent studies may be missing because the search engines may not have indexed them.

Finally, exclusion of papers written in a language other than English may have lead to bias in the selection process. This could not be avoided due to impossibility of the revision team to address these languages.

The second limitation concerns bias in the data collection. The first information collected was about the identification and general details of the paper such as title, its authors, source where it was published, abstract, year of publication. Full texts have been read when abstract of papers not provide enough information.

**Conclusion and future work**

This paper presents the results of a systematic review about the usage of techniques on software repositories until January of 2014. After the searching and filtering process, 169 papers were analyzed.

They have cataloged the papers selected based repository using the techniques of data mining that use storing spreadsheet data needed to answer questions research.

Analyzing results of this study note that interest in software repositories usage and trend to employ neural networks techniques to data mining is increasing.

In summary, this paper presents a comprehensive snapshot of actual use of neural network to analyze software repositories.

In the future, the authors intend to explain the period of gathering of papers to keep the work up-to-date and the answering other interesting questions concerning the trend of usage of neural networks. Other future work will delve into the use of neural networks as a tool for mining software repositories to facilitate the work of developers and project managers.

**Bibliography**


Appendix A

This section provides the primary studies selected from the systematic review


Krishnan, S. (2013). Evidence-based defect assessment and prediction for software product lines. PHDThesis


Litoriya, R., Sharma, N., & Kothari, A. (2012, September). Incorporating Cost driver substitution to improve the effort using Agile COCOMO II. In Software Engineering (CONSEG), 2012 CSI Sixth International Conference on (pp. 1-7). IEEE.


Minku, L. L., & Yao, X. (2012, June). Using unreliable data for creating more reliable online learners. In Neural Networks (IJCNN), The 2012 International Joint Conference on (pp. 1-8). IEEE.


Smith, M. R., & Martinez, T. (2011, July). Improving classification accuracy by identifying and removing instances that should be misclassified. In Neural Networks (IJCNN), The 2011 International Joint Conference on (pp. 2690-2697). IEEE.


Authors' Information

**Ana Maria Bautista** – E.T.S. Ingenieros Informáticos. Universidad Politécnica de Madrid, Madrid; e-mail: am.bautista@alumnos.upm.es

*Major Fields of Scientific Research: Artificial Intelligence*

**Tomás San Feliu** – E.T.S. Ingenieros Informáticos. Universidad Politécnica de Madrid, Madrid; e-mail: tomas.sanfeliu@upm.es

*Major Fields of Scientific Research: Software Engineering, Computer Science*