On Software Engineering Repositories and their Open Problems

Daniel Rodriguez, University of Alcala, Spain

Israel Herraiz, Technical University of Madrid, Spain

Rachel Harrison, Oxford Brookes University, UK

Universidad de Alcalá

OXFORD BROOKES UNIVERSITY

Realizing Artificial Intelligence Synergies in Software Engineering (RAISE), 5 June 2012

Outline

Introduction

Repositories

Classification of the repositories

Some open problems/issues

Conclusions and future work

SE Repositories – Information

Source Code

Source Code Management Systems (SCM)

Issue Tracking Systems

- Communication
- Mailing List, IRC, forums

Meta-data about the projects

- Other information such as programming languages, domain, licences, project management data (e.g., effort, personnel) defects, etc.

Usage data

- Number of downloads from the Internet, usage data

SE Repositories classification

Type of information

- Meta –information about the project and personnel
- Low-level information
  - Mailing Lists, IRC, forums
  - Bugs tracking Systems (BTS) or Project Tracking Systems (PTS)
  - Processed information (effort estimation, cost)

Whether the repository is Public or not

Single project vs multi-project

- Multiple versions of the same project or multi-project (and versions).

Nature of the project: Open source vs. Commercial project

Format of the information

- Text: plain text, CSV, ARFF – Weka's format
- SQL – Database dumps
- Remote access – Web services or REST

Some Repositories

FLOSSMole: http://flossmole.org/

FLOSSMetrics: http://flossmetrics.org/

PROMISE (PRedictOr Models In Software Engineering) http://promisedata.org/

Qualitas Corpus (QC): http://qualitascorpus.com/

Sourcerer Project: http://sourcerer.ics.uci.edu/

Ultimate Debian Database (UDD): http://udd.debian.org/

Bug Prediction Dataset (BPD): http://bug.inf.usi.ch/

International Software Benchmarking Standards Group (ISBSG): http://www.isbsg.org/

Eclipse Bug Data (EBD)
http://www.st.cs.uni-saarland.de/softevo/bug-data/eclipse/

Software-artifact Infrastructure Repository (SIR): http://sir.unl.edu/

Oohloh: http://www.ohloh.net/

SourceForge Research Data Archive (SRDA) http://zerlot.cse.nd.edu/

Helix Data Set: http://www.ict.swin.edu.au/research/projects/helix/

Tukutuku: http://www.metriq.biz/tukutuku/

SECOLD: http://www.secold.org/

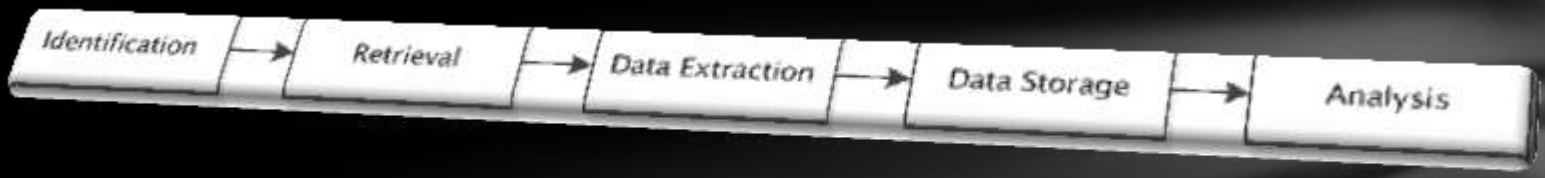
Comparison of Repositories

	Meta-info	*Single vs. Multi*	*Open?*	*Format*
FLOSSMoles	Y	Multi	Y	DB dumps, text, DB access
FLOSSMetrics	Code related	Multi	Y	DB dumps, Web srvs, Web
PROMISE	Some DSs about Proj. Manag	Multi	Y	Mainly ARFF, CSV, others
QC	Y	Multi	Y	Code, JAR, CSV
Sourcerer	N	Multi	Y	Java Source code + DB structre
UDD	Y	Y	Y	DB dumps
BPD	N	5 Eclipse proj	Y	CSV
ISBSG	Proj Manag data	Multi	N	MS Excel Spreadsheet
EDB	N	Eclipse	Y	ARFF and CSV (same info)
SIR	N – for testing	M	Y	Code for analysis and testing tech
ohloh	Y	Multi	Y	Web (limited)
SRDA	Y (SF.net)	Multi	Y	DB Dumps
Helix	Y	Multi	Y	CSV
Tukutuku	Proj Manag	Multi	N	Effort pred for Web apps

Extracting Information

Similar to the general data mining process by Fayyad et al.

- But it has its own characteristics and difficulties (Robles et al.):



Large variability in the formats and tools needed, standards, etc.

- Mining of textual data to deal with bugs for classification, clustering, find topics, etc.
- Regular expressions, Information Retrieval techniques, etc.
  - Difficult task even with human intervention because change requests and incident reports are often mixed together in the BTS or PTS.

Replicability

Replicability is one of the main reasons to adopt open repositories (Kitchenham et al.). However…

- …Risk of replicating experiments without using the original sources.
  - Preprocessing is the hardest tasks in the data mining process.
  - Trusting the preprocessed data from others can be a poisoned chalice.
  - For example, Shepperd has reported differences between using a original NASA datasets or a preprocessed one downloaded from the PROMISE repository.

Not many works provide the necessary means to replicate the studies

- Eclipse Bug Data contains the data and scripts to replicate the study.
- Robles et al from analysing MSR papers: ***"A total number of 171 papers have been analyzed […]. Results show that MSR authors use in general publicly available data sources, mainly from free software repositories, but that the amount of publicly available processed datasets is very low. Regarding tools and scripts, for a majority of papers we have not been able to find any tool, even for papers where the authors explicitly state that they have built one."***

Open Issues
Outliers, Missing Values and Inconsistencies

Outliers

- Although this statistical problem is well known in the literature, it is not always properly reported for example in many estimation studies as stated by Turhan et al.
- Seo and Bae (ESE 2012) - Effort prediction with and without outlier elimination differs depending on the dataset used.

Missing values and inconsistencies

- Some of the repositories such as the ISBSG, are composed of a large number of attributes, however, many of those attributes are mainly missing values
  - that need to be discarded in order to apply machine learning algorithms.
  - Or use imputation methods
- There are also inconsistencies in the way information is stored. In this particular dataset, cleaning inconsistencies (e.g., languages classified as 3GL or 4GL, Cobol 2 and Cobol II, etc.).

Open Issues
Redundant and irrelevant attributes and instances

Irrelevant and redundant features in the datasets has a negative impact in most data mining algorithms.

Feature Selection and Feature Ranking have been applied and studied by the software engineering community,

- not so much instance selection which needs further research (a few exceptions for effort estimation include Chen and Menzies, IEEE SW).
  - E.g. JM1 from the PROMISE repository has around 8,000 repeated row

It is known, however, that feature selection algorithms do not perform well with imbalanced datasets

- resulting in a selection of metrics that are not adequate for the learning algorithms. This problem can happen in most effort estimation or defect prediction datasets as mentioned before such as the ISBSG that has over 60 attributes most of them are irrelevant.
- Also the defect prediction datasets such the EB data are highly unbalanced. Some further research into robust algorithms such as Subgroup Discovery techniques is also needed [28] or weighting of attributes and instances.

Open Issues – Overlapping

When dealing with classification, we may also face the problem of overlapping between classes in which a region of the data space contains samples from different values for the class.

We have found that many samples from the NASA dataset contained in the PROMISE repository are contradictory or inconsistent, many instances have the same values for all attributes with the exception of the class, making the induction of good predictive models difficult.

Open Issues – Data shifting

The data shift problem happens when the test data distribution differs from the training distribution. Turhan discusses the *dataset shift* problem in software engineering (effort estimation and defect prediction).

It is customary in data mining, to preform the evaluation using cross-validation, i.e., divide the dataset into *k*-folds for training and testing and report the averages of the k folds. This problem can easy happen when we are dealing with small datasets.

Also when we are dealing with small datasets, it can happen that the number of instances that remain in the training dataset is skewed. Many software effort estimation datasets are very small (around 20 effort estimation datasets contained in PROMISE repository contain just over a dozen samples, e.g., the Kemerer or Telecom datasets)

Open Issues – Imbalance

This happens when samples of some classes vastly outnumber the cases of other classes.

In this situation, many learning algorithms generate distorted models for which:

- the impact of some factors can be hidden
- the prediction accuracy can be misleading

Although a well-known problem in the data mining community, this problem has not been addressed in detail by the SE community. Typically addressed by:

- pre-processing the datasets with sampling techniques
- or considering cost in the data mining

This problem happens in many of the defect prediction datasets (e.g. the PROMISE repository has around 60 defect prediction datasets).

The previous problems, redundant and irrelevant attributes, overlapping, data shifting and small datasets are made worse when datasets are imbalanced.

Open Issues
Metrics and *fitness functions*

In relation to the measurements, either from the social network data, mailing lists or code, there can be differences depending on the tools used in those repositories that contain source code such FLOSSMetrics, EBD, or BPD.

For example, Lincke et al. report on large differences in metrics collected from the code depending on the tool used.

Shepperd and MacDonell report on the abuse of using MMRE (Mean Magnitude of Relative Error) when dealing with effort estimation.

- MMRE has been known to be biased and favours underestimation, perhaps because it is easy to apply, it has been used to wrongly validate and compare different estimation methods or models.
- Such metrics can be used as fitness functions in metaheuristic algorithms (Harman and Clark, Metrics'04), the solutions obtained may be suboptimal.

Conclusions and Future Work

Number of repositories is increasing, mainly thanks to open source

Large empirical studies,

- From the statistical and machine learning poing of view
- Closely related to SBSE

Future work

Analysis of Open source data mining tools

and their adaptation to SE problems