Understanding and Prdicting Method-level Source Code Changes Using Commit History Data

by

Joseph Heron

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science

in

Computer Science

University of Ontario Institute of Technology

Supervisor: Dr. Jeremy Bradbury

April 2016

Abstract

Software development and software maintenance require a large amount of source code changes to be made to a software repositories. Any change to a repository can introduce new resource needs which will cost more time and money to the repository owners. Therefore it is useful to predict future code changes in an effort to help determine and allocate resources. We are proposing a technique that will predict whether elements within a repository will change in the near future given the development history of the repository. The development history is collected from source code management tools such as GitHub and stored local in a PostgreSQL. The predictions are developed using the machine learning approaches Support Vector Machine and Random Forest. Furthermore, we will investigate what factors have the most impact on the performance of predicting using either Support Vector Machines or Random Forest with future code changes using commit history. Visualizations were used as part of the approach to gain a deeper understanding of each repository prior to making predictions. To validate the results we analyzed open source Java software repositories including; acra, storm, fresco, dagger, and deeplearning4j.

Acknowledgements

Contents

List of Figures

List of Tables

Abbreviations

ANN Artifical Neural Network.

API Application Programming Interface.

DVCS Distributed Version Control System.

IR Information Retrieval.

LD Levenshtein Distance.

MSR Mining Software Repositories.

NLD Normalized Levenshtein Distance.

OS Oversampling.

OSS Open Source Software.

RF Random Forest.

SQL Structured Query Language.

SVM Support Vector Machine.

SVN Apache Subversion.

SWR Sample Window Range.

VCM Version Control Management.

VCS Version Control System.

Chapter 1

Introduction

Software has become pervasive and integrated with numerous platforms and applications such as mobile devices, web sites, embedded systems, safety critical systems. Creating and managing a software application can be time consuming and resource intensive. The development of software applications commonly integrate the usage of Version Control System (VCS) to manage the application by storing the current version as well as previous versions in a repository. The development of a project is limited by the resources available to the team developing the application. Effective allocation of these limited resources could be the deciding factor in whether project will be successful or not.

1.1 Objective & Methodology

The mining of open source software repositories is widely used to help research into various software topics relating to software development and quality assurance. Research can provide improvements to the development process of software repositories. With an improved development process, more repositories may succeed in accomplish-

ing their outlined goal. The process of developing a repository will of course take time to complete. The time for a repository to be completed relies on numerous factors including scope, man power, experience. Over the course of development, changes will be made to repository. Changes can be made to almost any part of the repository including design, number of developers and type of developers. These changes will in most cases have a measurable impact on the repository. In case of adding more developers, the intended result may be to increase functional capabilities within a shorter span of time than previously. Even with an intended result, the actual result may differ and should be measured to determine the effectiveness of a given change.

The developers of the repository should manage the growth of the repository to ensure that the changes that are made result in an expected outcome. Keeping track of every change to a repository can be difficult because of external changes which are beyond the control of the developers. However, for the majority of the changes within the repository they are kept track through VCS. With proper use of a VCS, the important changes made to the repository availableb. This can help keep previous releases of the software available or even help resolve a bug that was introduced in a recent change. Furthermore, developers have control over what is stored in the VCS allowing for granularity based on developer preferences. With numerous developers, a VCS can also help improve how these developers interact and share the changes they made. Some commonly used VCS include Git [1], Apache Subversion (SVN)[2] and Mercurial[3].

The impact of changes can be measured and provides insights into how the repository changes. However first the data must be collected, processed and stored. While changes may occur in various forms, a more accessible one would be the source code

[1] `https://git-scm.com/`
[2] `https://subversion.apache.org/`
[3] `https://www.mercurial-scm.org/`

changes in the software repository. These changes are very fine grain since they will account for almost all functionality changes with the repository. The only functionality changes not accounted by source code would be external changes (e.g. library changes). Changes will map to functionality changes that provide fixes, new functionality, or removal of functionality. The source code changes will provide a large amount of noise since every change is included. This excessive granularity can make tracking the desired changes more difficult. Visualization of the data collected allows for a more accessible look at the data to provide potential insights.

As discussed earlier, there are two main types of software repositories that are developed, either closed source or open source. Open Source Software (OSS) repositories will generally provide access to the source code, the ability to change and finally redistribute the changes. OSS is widely used in developing software repositories of various sizes and scope. In these repositories developers are able to contribute towards the a repository that is often used by a wider audience. While smaller OSS repositories may have a small number of developers, larger repositories can contain developers from numerous locations around the world contributing at different times. The development of OSS is often the focus of research related to software development since the repositories are open and freely available. The authors are able to publish and use the data as they wish since it is publicly available. There are also countless OSS repositories to study and investigate.

The collection of data is done through data mining. Data mining is the act of collecting data from one or more sources to use for another goal. Often data mining will use a data source not traditionally used, since the goal of data mining is to extract and use information. The actual use of the data once collected can vary greatly from visualizing to modeling. Data can also be collected in several forms including continuous streams of data, sporadic data and one time collection. Depending on

what type of data is being collected and the purpose of the collection the means of collection may also vary. Another concern related to data mining is that of Big Data. If a source provides a wealth of data, then extra measures should be taken to manage the size of the data set. Without diligent management, a data set can become unwieldy with massive overhead that are entirely avoidable.

The goal of the approach is to predict changes that will occur within the repository using the commit history. In this case machine learning techniques are leveraged to create a model based on the data collected through mining GitHub to predict changes. Machine learning techniques are widely used to support the completion of difficult tasks that involve patterns. A machine learning algorithm is generally an algorithm that attempts to detect and mimic patterns within a data set. There are numerous different types machine learning algorithms including SVM, RF, Artifical Neural Network (ANN). Each technique provides advantages and disadvantages depending on the purpose and the data set in use. The primary focus will be on SVM and RF since they are used as part of the proposed work. In order to create the predictive model, a subset from the data is used for the training of the model. To test the model a second subset that is distinct from the first is necessary to allow measure the performance of the model.

SVM is an algorithm that attempts to classify data into two distinct categories. This algorithm is a supervised learning technique that requires a training data set to build the categorization model. The training set will consist of data samples from each classification as well as which classification the data sample belongs to. After creating the model for a SVM new data vectors can be provided to the model and be classified into one of the two categories. The model will be constructed by attempting to linearly separate the data into two distinct groups. If the data cannot be separated linearly, then the data is mapped to a higher dimension to allow for proper separation.

During the separation of data points into two groups, the model may reclassify data points in an attempt to fix errors within the data set. This feature allows for some error to be present within the training set without causing further errors. In the case that points which are valid are detected as errors then the separation has generated errors and the features or data used may not be useful towards making a prediction.

RF is another supervised learning technique that requires a training data set to create an prediction model. The origin of random forests is the decision tree learning method. A single decision tree creates a tree structure where each internal node in the tree represents a decision where in the final destination is the outcome. RF extends decision trees to address the tendency for a decision tree to overfit the data. A RF uses numerous decision trees as well as a modified version of bootstrap aggregation to get more robust predictions.

The machine learning algorithm makes use of the training data to create a model for predicting the classification of a given value. An analysis of change data can help in the selection of useful information for creating the predictive model. The data is collected from the project's historical data that can provide a large data set to work with. Managing the data and selecting ideal samples for training must be considered to provide a strong predictive model.

We propose a tool that assists in managing the development of software repositories by predicting changes that are likely to occur. This work explores leveraging change prediction of the source code using the commit history to assist in the development of OSS scale repositories. The key factors; sampling size, feature set and data balancing are investigated using the tool to provide a deeper understanding the feasibility of the tool. Several OSS repositories were selected to conduct experiments to determine the impact of each of the factors.

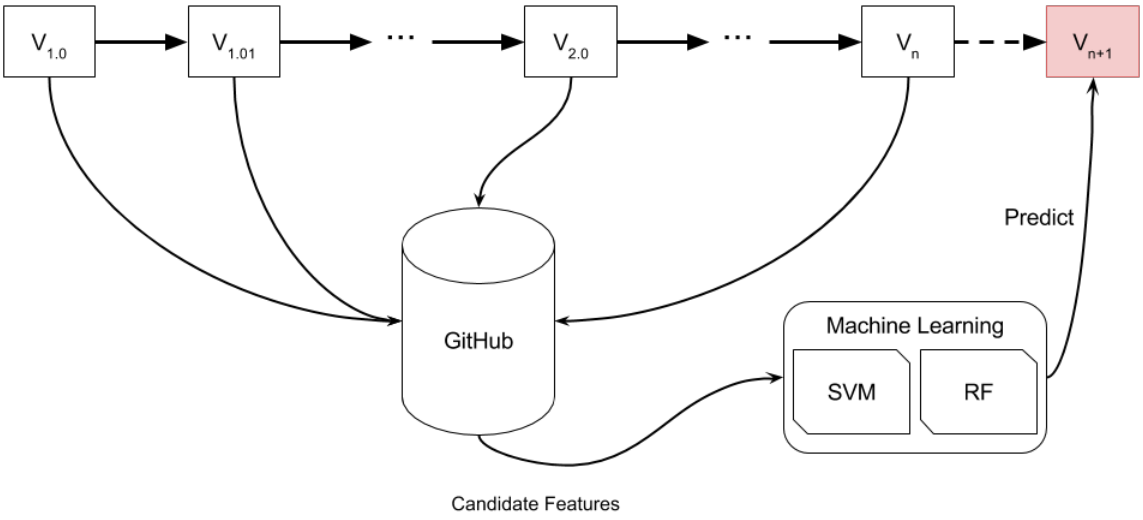


Figure 1.1: Approach Overview

1.2 Contributions

Our contributions are in mining of OSS, visualization of a repository's change history, machine learning change prediction, data collect which can be used and extended. Providing clear and accessible visualizations allows for the development data to be inspected more thoroughly. Likewise, strong predictions of change can assist in the development of more robust, efficient and less costly software programs. Finally, the collection of historical data for use of predicting future changes is presented as a possible option when predicting future changes within a repository.

There are several different areas where this work can be applied and provide improvements. The main area which this research is applicable would be that of software development by providing support during the development process. The prediction of future changes within a repository are made available to developers to support them when choosing tasks or making new changes to the repository. With

the knowledge of where changes are likely to occur within the repository developers may be more prepared in making such changes. This is especially true if this work was extended to provide more specific change information about future changes. Another potential use for this approach would be in resource allocation for software repository development. A larger repository with numerous developers contributing will require each developer to work on various task an attempt to limit conflicting contributions. With the ability to predict where future changes will likely occur developers can coordinate more effectively with other developers mange interactions and overlap. Finally, the approach could help with resource allocation for a VCS provider, such as GitHub. A VCS could use the repository data to efficiently allocate resources based change predictions for each repository. If a repository is likely of have near future changes then more resources are necessary for that repository compared to one that is less likely to receive changes in the near future. With strong predictions a system could be more effectively managed and offer savings to the company.

1.3 Organization

The remainder of the thesis is organized into 5 more chapters.

1. Literature Review which provides more details related to the foundation of this work. Primarily this chapter will cover the data that is collected for the analysis.

2. Visualization of Commit Data discusses how the data is collected, stored and visualized.

3. Prediction of Commit Data outlines the data and methods that are used to predict change within a repository.

4. Experiments reports the experiments conducted and their results.

5. Conclusion summarizes the results and contributions and proposes future work to build of the thesis.

Chapter 2

Literature Review

2.1 Data Mining

Data collection from some original source provides access to a data set that may not be initially available. This data source could also be in a state that is not convenient or feasible for use without leveraging data mining techniques to transform the data to a more accessible state. The source of the data can vary greatly based on the interests for the individual(s) collecting the data. Data mining has mostly focused on single source mining and multiple data sources. Data mining in general has however also taken a large focus on data collection from software repositories which can be either single or multiple source [10, 19, 21, 24, 29, 44].

Zimmermann et al. collect change the version history of a software project to predict changes that should be made in relation to an initial set of changes. The recommendations their tool provides helps point the developer to make changes that are more common within the project. As well the tool can be used to detect which changes may be missed by a developer when making changes to a project. Maletic and Collard investigate source code changes during a software project's development

cycle. The changes are extracted and stored in an more easily usable form to be more easily analyzed. Canfora et al. propose a method for extracting and refining the changes made throughout the life a project to be used in more effective analyses. The changes made to a project are refined through linking lines of source code that are related. Hemmati et al. take a comprehensive review at the research related to Mining Software Repositories (MSR). Several best practices are proposed and areas of future work are identified. Hassan discusses the value of data mining from software repositories. The possible uses of the data collected can be used towards are assisting developers or managers. A benchmark data set of software project development change history is provided by Dit et al. The data set is processed to provide change request description and tracing, where changes that are requested are able to be traced to where they were implemented within the source code. The data set also provides a corpus of various key aspects of the project including files, classes and methods. The data set is targeted to be used for providing a benchmark for tools attempting to improve software maintenance tasks.

2.1.1 Mining Open Source Software Repositories

OSS generally is software that provides with the ability access the source code and make modifications to the source code. While certain licenses provide some restrictions on the ability to redistribute the software the main point of the source code of the software being freely available is key. The scope and capability of OSS projects vary greatly. Several very popular OSS projects are listed in Table 2.1.

The development of large software projects (whether OSS or not) often make use of VCS. A VCS helps the developers of the project manage the changes of the project and facilitate the collaboration between developers. A VCS will keep an current version of the project and keep track of the previous version of the project as well.

Owner	Project	Description
Mozilla	Firefox[a]	Internet Browser
Linux	Linux Kernel[b]	Operation System Kernel
VideoLAN	VLC[c]	Media Player
PostgreSQL	PostgreSQL[d]	Object-Relational Database Management System
git	git[e]	Version Control System

Table 2.1: Open Source Software Projects

[a] https://www.mozilla.org/en-US/firefox/desktop/
[b] https://www.kernel.org/
[c] http://www.videolan.org/vlc/index.html
[d] http://www.postgresql.org/
[e] https://git-scm.com/

This may be done through keeping a copy of each version of the project or by keeping track of all each change made to the project. SVN and git would be two examples of VCSs.

Git is a Distributed Version Control System (DVCS) and differs greatly from SVN which is a normal VCS. Git will provide the user with a complete copy of the repository that is worked on independent of network connection. The independence of each repository also allows for a repository to be developed without a centralized server. The distributed aspect of git tends to allows for easier use for all involved parities. The one main issue with a DVCS is that while decentralization is useful, developers will require some method to collaborate and communicate to transfer changes made to the repository. Therefore typically one centralized server is used to maintain communication between all interested parties.

Git has grown in popularity since it was created and is at the core of several Version Control Management (VCM) sites such as GitHub [1], BitBucket [2] and GitLab [3]. These platforms tend to be fairly supportive of OSS projects through providing their

[1] https://github.com/
[2] https://bitbucket.org/
[3] https://gitlab.com/

services free of charge. For example, GitHub provides unlimited public repositories completely free. While these projects do not have to be licensed with an open source license typically they will be since they are already publicly visible.

GitHub is the most popular of the VCM websites and hosts numerous very popular OSS projects including, the Linus Kernel, Swift[4] and React[5]. GitHub also provides a public Application Programming Interface (API) to allow for access to the data related to project repositories which is discussed further below. Given the popularity of GitHub for use by developers and the availability of the project data, GitHub is an obvious choice for mining project data. Especially since the goal of mining software is to capture OSS project data to both explore and test analysis methods. Publicly visible projects are also publicly accessible through the API and the majority are open source.

Git provides a simple interface to manage the repository regardless of which site is the central server. Therefore regardless which site the project resides on users can easily interact with the project as long as they know the git interface. Git in essences is a file storage for the project that keeps track of changes made to the project. A *commit* is a set of changes that a developer has made at a certain time. The developer has full control what gets committed, when it gets committed and even modified at a later date.

A branch is a series of commits that are often related. In Figure 2.1, each dot would represent a commit and a set of dots connected by the same colored lines are a branch. Branches can be considered different paths or deviations in the development from each other allowing for different versions of the project to be maintained and developed. The *master* branch is the main branch, represented with black, from

[4] `https://swift.org/`

[5] `https://facebook.github.io/react/`

which all branches usually stem from and is generally where projects are developed on. On a similar note, a *tag* is a branch that is frozen to allow for future reference. Tags are often uses to mark a significant point in the development history such as a project release. Finally, when two differently branches converge into a single dot then the two branches have been *merged*. A merge indicates that the differences between the two branches are consolidated based on the developer's discretion.

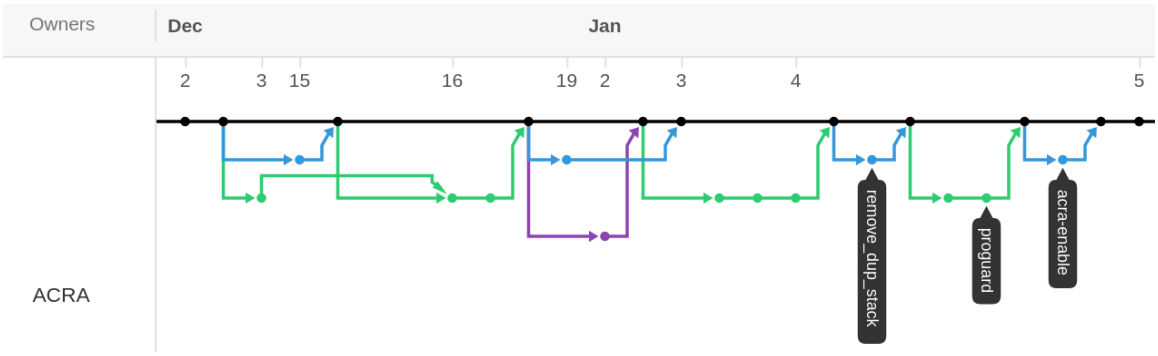


Figure 2.1: Network diagrams

A commit consists of files that have been changed, more specifically a list of *patch* files which each outline the changes made to their corresponding file. The patch file consists of a series of differences between the previous version of the file and this new version of the file. These patch files are key since they contain the actual changes made to the project and thus are the major point of interest.

2.1.2 Visualization

Visualizations are often used to represent data to be more accessible to appealing for use. Rather than view large amounts of complex data a visualization can restrict the amount of information shown to prevent the user from being overwhelmed. Alternatively, a well designed visual representation of the data can retain information and represent it in a way that is more convenient. Visualizations are widely used

throughout research to represent software evolution [5, 8, 12, 16, 25, 36] and developer interactions [9, 14, 16, 36, 37].

Some of the visualizations attempt to focus on a particular aspect. However, by their very nature of evolution visualization the visualization attempts to showcase higher level information. Gall and Lanza discuss uses of project traits including the source code changes, release information and quality metrics to provide the necessary data for powerful visualizations. Similarly, Collberg, Kobourov, Nagra, Pitts and Wampler use CSV version control systems to visualize software evolution. A visualization is produced which provides a temporal element for development data. Another approach was to study visualize the change-proneness within the software project [5]. Lanza, Ducasse, Gall and Pinzger present a high level visualization tool for object-oriented projects. Ogawa and Ma create a story view visualization for software projects with an summary of changes for each commit. The visualization can have issues with large amount of information available.

Ogawa and Ma provide a expansive visualization which includes developer and source code interactions. Gonzalez, Theron, Telea and Garcia visualize a combination of software project metric and structural changes. Four designs are proposed to provide unique and complementary views.

2.2 Machine Learning

Machine learning is a complex method for software algorithms to attempt to determine patterns within the data. One such problem example would be an algorithm to detect certain people within an image. For an individual such a task may seem trivial however for a software system to detect it is far more difficult. Algorithms that can determine patterns and mimic them from abstract set of data is useful when such pat-

terns are extremely complex. There are numerous algorithms which apply machine learning approaches. Each approach has both advantages or disadvantages. Some examples of machine learning algorithms are SVM, RF and ANN. The three provided examples are also commonly used for data mining [1, 4, 17, 22, 23, 42]. Bhattacharyya et al. provide a detailed description of RF and SVM.

2.2.1 Support Vector Machines

A SVM is used to predict what type of change will occur based on a set of features provided. A feature is a data extracted from the project represented as a floating point number. In order to be useful a feature must in some way characterize the the category that it is assigned to. The feature must also not rely on the category that it belongs to in order to be calculated. For example, given a category of the method change within the next 5 commits or not, then the features must not rely on knowledge of future changes to the project. If the features fail to effectively characterize the category they are assigned to then the SVM may have poor predictions. It is also necessary for the features to independent of each other to not negatively affect the categorization.

SVM has been widely used for making predictions for various aspects including predicting battery charge state [2], pharmaceutical data [6], software faults [11, 15, 28, 30, 31, 35], bug localization [33, 35], software mutation testing score [23], financial stocks [27], credit score [22], credit card fraud [4], solar power output [43].

Malhotra reviews numerous machine learning techniques, including SVM and RF, used by various studies. The results of which outline where each approaches succeed and falls short. When using a machine learning algorithm it is imperative to use a suitable algorithm for current situation. Kim et al. outline a approach that uses a SVM to predict changes that will occur within the project. By identifying these changes the a project developer can potentially locate a bug within a change and fix it

prior to being reported. Erturk and Sezer compare the performance of their proposed method, an Adaptive Neuro Fuzzy Inference System, to that of an SVM for predicting software faults. The models are trained using project metrics as well as the project's historical fault data. Zeng and Qiao use a SVM to provide short-term predict solar power output. The SVM model outperformed both an autoregressive and a neural network model. Anton et al. propose a method for predicting the state of charge of a battery using SVM model. Neuhaus et al. mines vulnerability databases and version archives determine components within the software that were vulnerable. A SVM was then used to predict other component that were also vulnerable. Several feature selection techniques have been assessed by Shivaji et al. for bug prediction methods. Features which are less useful to the prediction are removed to reduce the set to only the essential features. Kim investigates the possible use of SVM as a prediction model for financial forecasting. The model was used to predict whether the stock price would go up or down for the next day.

Bhattacharyya et al. uses RF, SVM, logistic regression to detect credit card fraud. Both RF and SVM are able to predict a large number of fraudulent credit card transactions.

SVM requires all feature data be encoded as floating point numbers. For any numerical data the conversion to floating point is trivial. However, for more complex data the conversion is a little more difficult. Categorical data can be mapped into a unique vector entry per category. For example, if a feature can be 1 of 3 options: 0, 1 or 2 then it can be converted into three entries in the feature vector. Encoding the value 2 the sub-vector of the feature set would be {0, 0, 1} where 1 indicates a field that feature is present in the data for this vector, and 0 indicates the feature is not present. Data that is in the form of a string can be converted to a floating point number by assigning a unique number for each string (similar to hashing). The one

downside to this method is that the numbers corresponding to each string maintain no numerical properties. In essence the data becomes categorical, such that if *bob* is mapped to 1 and *sally* is mapped to 2 there is no relationship between 1 and 2. Ideally, this data would then be further converted using the previously described method however if the set of possible strings is large then it may be unreasonable to convert it. For example, if there are 100 possible strings then that would add 100 new entries to a single vector.

The categorization is used for the prediction, where each value of the category relates to a unique prediction type. For example, a simple binary categorization could simply 1 or 0 where 1 predicts the event will occur and 0 predicts that the event will not occur. In essence an SVM is tasked with separating a dataset into two different categories given a sample set of data that has already been categorized into two subsets. Given the categorization of the sample dataset the SVM model is trained to allow for categorization of new data. The categorization of any new vectors (that were not used for training) is called a prediction and is made by the SVM model created through the training. More specifically, the sample dataset is a dataset extracted from the target dataset. The sample dataset is then categorized based on the predetermined criteria (the prediction goal). This dataset along with the categorization for each vector in the dataset is the training dataset, and is then used to *train* the SVM model. Once the model has been trained, the SVM model is ready to be used for making classification predictions. The data for each feature can be extracted from the new dataset, allowing for the model to classify each new vector. Given that the SVM model is accurate and reliable the results can then be used towards making predictions about the dataset. For example if the classification is that of predicting change to occur within the next six commits the developer may wish to be careful with the use of the method or assess the method's quality and

determine if any issues within the method need to be addressed.

A lower prediction score often relates to the data from the feature set poorly characterizing the categories. Similarly a warning will be given if the dataset is inseparate. In this case, the dataset for each category may be too similar and cannot be properly split into the two category subsets. In both cases a change to the feature set may help, whether that is a decrease or increase of features in the set. Some features are detrimental to the model, especially two features related to one another.

More details about the specific features used will be given a little later on. Features are descriptive aspects of the dataset that are classified into the predetermined categories. Since these features relate directly to the category understanding of the classification critical and can help determine which features should be used. For example for a classification of whether a change will occur within the next few commits, a useful feature may be the frequency by which a method changes within the project. Picking a descriptive feature set is paramount to providing a strong prediction of future data.

Most of this was done using database queries or user defined functions created in the database language.

2.2.2 Random Forests

RF are a popular machine learning algorithm and is used in numerous areas including predictions for software fault [18, 30, 31], software development effort [31], credit card fraud [4], database indexing [42], malware detection [1].

Malhotra provides an extensive review of studies involving machine learning to predict software faults. The results showed that RF tended to preform better than other machine learning algorithms studied. Moeyersoms et al. made use of RF and SVM as well as a few other data mining approaches to predict software faults and

effort estimation. The data mining techniques are used as part of another model, ALPA rule extraction, to improve the predictions and increase traceability. Guo et al. attempt use RF to predict the fault proneness of modules within a project. The RF prediction results for the five sample projects prove more accurate to that of a logistic regression. Yu et al. attempt to use RF to determine a more effective database indexing for video data. The database index are used to provide faster searching of the database for action detection.

RFs are commonly used to on data that has been mined from some source to make predictions [1], [17], [42]. A RF leverages numerous decision trees to provide attempt to improve prediction capabilities. Therefore to fully understand a RF first an understanding of decision trees is necessary. A decision tree is a technique which will create a tree based on a data set that has been classified. Once the decision tree model is created it can be used to predict or categorize data that has not yet to be classified. In the tree model the leafs will be categorizations where as the connections between inner nodes are the decisions by which the categorizations are made.

One issue with decisions trees and more generally machine learning techniques in general is imbalanced data sets for training the model [26]. The data set used rarely provided even sample sizes of each set therefore without taking necessary pro-cautions the algorithm will bias the results. In the worse case the model will classify any input data as the larger data classification.

In case of imbalanced datasets there are several methods to help provide stronger predictions [26]. The most obvious and easiest to attempt would be to sample more data. However if the dataset in general follows this trend then some more advanced techniques can used to improve the model.

The first method would be to *undersample* larger category this will even out both of the categories. This will remove some of the input values within the dataset to

reduce the set size. However if there are very few samples of the smaller category the performance will suffer as well. A second method of *OS* is useful in the case were the data samples are small. The input data from the smaller category is selected to be duplicated in the set to increase the size of the set. This helpful since it will increase the size of the dataset but could lead to bias based on the data selected from the smaller dataset. The selection method for which input vectors to over or under sample can be based off on the data's statistical distribution or made by random choice. Another advantage of these over and under sampling is that they can also be used together to in the case of a large disparity between the category's set size.

Another feature of RF which is used to help provide more reliable predictions is *Bootstrap Aggregation* [4]. Similar to normal sampling methods it will take the initial dataset. However rather than using the dataset as is the dataset will be uniformly sampled n times and repeated m times to create m datasets of n values. These newly created datasets will then be used to train m models. Finally, when attempting to categorize a new input data it will be given to every model and the prediction result will be aggregated to provide a more accurate results. For some machine learning methods such as SVM this method will improve the results and help with imbalanced datasets.

A RF is a collection of decisions trees trained on random samples of the initial dataset. So the RF will take an input dataset and then train m decisions trees using m randomly sampled sub-datasets of the initial dataset. This helps improve the model created and makes RFs far easier to use. As well RFs have a feature that determines the importance of each feature is assessed during the training of the model [4]. The importance outlines the quality of each feature in providing the prediction [40]. Therefore in order to properly understand the feature importance the accuracy, precision and recall of the model should be determined by running a test

dataset to determine the quality of the model.

2.3 Software Development Prediction

The development of large scale projects can take a long time and involve a huge time investment from the developers. The development of the project will cause for the developers to make changes to projects. Changes made to a project may introduce new faults, increase functionality and fix previous problems. Therefore changes to a project can be both positive or negative. The developers of a project must control how a project is changed to attempt to limit the number of negative changes and increase the positive changes. Beyond ensuring that the project is developed correctly the developers typically have a limited amount of time to spend on the project and therefore must allocate their time wisely. Software development prediction models are used to help developers allocate their time more effectively. For example a developer may have a list of features that should be added to the project. However implementing the most fundamental features first will help ensure that these features are more likely to be completed.

Software development prediction contains numerous areas of study which generally attempt to improve projects by focusing on their development and providing feedback to the developers through predictions. Some of these areas include predicting: fault detection [32, 34, 38, 39], mutation score [23], software changes [3, 7, 13, 20, 24, 41]. While there may be a large overlap in the objective for these studies often they will vary in what is used to make the prediction.

2.3.1 Fault Prediction

Fault prediction is a key area of study for software development since the goal is to provide insight into where issues within the project are located. Identifying these areas can be very beneficial to the developers in saving time from searching for bugs. Rather the developers are able to use their time on fixing those issues. Therefore accurate identification of faulty code improves both development efficiency and software product quality. In order to predict these faults studies used one or more of the following; change metrics [32,34,38], code metrics [32,39], defect history [38], software dependencies [34].

Fault predictions using static and change metrics are studied by Moser et al. The change metrics used outperformed the static metrics in accuracy, and recall. Sisman and Kak alternatively look specifically at change metrics using Information Retrieval (IR) framework to provide the predictions. The prediction framework also uses a time sensitive factor to bias towards more recent changes for predictions. Nagappan and Ball attempted to predict post release project failures for commercial projects. These predictions were done using a software dependency analysis as well as using churn metrics from the project's development. Their method proved to be capable in predicting these failures providing an ability to mitigate these failures from occurring.

2.3.2 Change Prediction

Software projects will have faults within the project especially during the development phase. A project in its early stages may not meet the full set of functionality since it has not been completed yet. Since the development team will known that such features are not yet implemented these faults or fails are not a huge concern. Rather faults that are unknown to the developer team are far more serious. Such cases as a

feature was thought to be implemented correct but was not or a feature implementation breaks other features. In both those cases changes made to the project cause the fault to be revealed. Changes to the project are the means by which all development occurs. The ability to analyze and predict changes within a project could give deep insights into the development of a project. A large amount of research as focused on predictions of changes based on changes [3, 7, 13, 20, 24, 41].

Ying et al. present a method that predicts which parts of the system will change given a set of changes or change propagation. The prediction is done using the project's change history. The results of the prediction method were mixed with some projects recording a stronger precision and recall and others recording a far lower results. Kagdi and Maletic also leverage version history changes to perform software change predictions. The actually analysis applied is two fold, through the dependency analysis of the current version and the change analysis of the version history. The data is collected through MSR which is a popular field of study. In a similar work, Hassan and Holt, worked towards predicting change propagation of a given initial change. The main question was to determine given a change to an entity (e.g. function or variable) will propagate to changes in other entities. This work is very related since it tests various methods and leverages presents the best one. Bantelay et al. propose a method that mines the file and method level evolutionary couplings to attempt to predict commits and other interactions within the project. Both methods were used in isolation as well to determine whether the attributes were more helpful when used together. Giger et al. attempt to build off of previous work in change proneness by providing predictions relating to more refined entities. While typical change analysis will involve the use of syntactic changes. However Giger et al. suggest that extracting and tracking semantic change could prove to be more helpful and accessible for developers for predicting future changes within a project.

Chaturvedi et al. attempt to predict the complexity of code changes to a project. The project's change history is analyzed and the entropy is calculated. The future amount of changes necessary, the complexity of code changes, is then predicted.

2.4 Change Analysis

Changes that occur within a project are made to achieve a goal or task. Whether the task is high level such as implement a new features for the program or lower level like fix a syntactical bug. Investigations into how changes are made or used can help provide a better understanding for making a better changes or better use of the changes.

Bieman et al. study the change-proneness of different entities within a software project. In order to provide a deeper understanding visualizations were used as well providing a bit of a different approach from some of the other works. Koru and Liu study and describe change-prone classes found within open source projects. Providing further details into characteristics of different changes that are made to a software project throughout development. Similarly Wilkerson attempts to classify different types of changes that occur to a project throughout development. The classification can then be used to identify the impact that a given change will have on other aspects of the project. Snipes et al. provide a tool that attempts to locate areas within the source code that have a large amount of changes. These areas could be classified as underdevelopment and are likely to be very unstable given the amount of change occurring within them.

Chapter 3

Visualization with Commit Data

The goal of the research proposes and assess a research tool for predicting changes within a project. This is accomplished through mining of software data, analysis of collected data, candidate feature analysis. Once the data has been collected a further analysis is used to extract key features. As part of this analysis, custom visualizations are used to help to provide insights into the data set. Candidate features are then selected from possible features and analyzed later on to attempt to determine the best feature set.

3.1 Collection

In order to be able to predict changes within a project, some project data must first be collected. The data collection is targeted towards OSS projects that are developed using GitHub. Specifically projects were selected that are predominately written in Java. A project would be predominately written in Java if it has over 75% of the source code in Java. Some of the sampled projects, especially larger ones, included other languages for small purposes such as a database schema outline. The purposed

approach is not language specific in theory, however in order to simplify the implementation the method was restricted to only work with Java. Other languages that will likely work are languages that are modular based around method or functions. Only a Java implementation was created, therefore the ability for other languages to be used with the approach may require redesign expecially prediction model features. The data collection process simply data mines the complete development history of the project through the commits stored in GitHub. The commit data includes developers related, the source files and the changes associated with them. Finally the project's release information is collected in the form of tags is recorded.

The data is kept unprocessed and stored directly into a relational database (MySQL) that allows the data to be used and manipulated without requiring access to GitHub again. This was ideal during the more initial phase of the research since a decoupled collection allows for various methods of analysis to be applied on the dataset without requiring the data to be download again. The collection process can take long to perform and depends largely on the size of the project. In the case of an incomplete collection, the collection process can be resumed to collect the remaining data. Similarly, a project that was previously collected can be mined a second time to collect any new commits made to the project. These maintenance collections will often be much smaller and require a smaller amount of time to collect.

The collection method chosen for mining data from GitHub projects was using GitHub's web API. The GitHub API allows access to the complete set of publicly available information stored in GitHub. Accessing the data through the web API allows for the collection process to be automated and vastly simplifies the process. This project commit history dataset can be rather large since it includes a snapshot of the commit, all the change data and developer data related. Therefore the process may take both a long time and lots of space. The collection process requires both

the name of the developer and repository. To actually collect the data from GitHub a Ruby script was used. This collection is built around a Ruby library, *github_api*[1], which is a convenient wrapper for GitHub's web API. The script systematically collects the desired data related from a given GitHub project which is stored locally. As noted above, the collection can take a bit of time to complete since it must go commit by commit to collect the necessary data.

Some aspects of the GitHub project's dataset are not collected as they were deemed unnecessary however the collection method could easily be extended to collect the other aspects. The aspects not collected are the issues, branches, forks and pull requests. The issues data outlines the problems reported in the project by users or developers of that project. GitHub allows for issues to be optional and thus some projects do not offer issue reporting through GitHub. Branches are also directly related to the project and they are essentially different workspaces for the developers. They allow for development of different versions (such as a development version compared to a stable version). For simplicities sake, the approach assumes that the main branch (the master branch) is the development branch and the target of the analysis. Therefore the branches are ignored at least for the scope of this work and could be included in future work. Of course other branches could be analyzed however the perspective of the other branches typically originates from the master branch.

A similar subset of data not collected or used for this approach is any forks of the repository. In GitHub a fork is an externally created branch of the project. The major differences between a fork and a branch are that a fork is owned by another developer and a fork is in fact a project onto itself. This allows for a developer who are not contributors to make a copy of the project and work on it without affecting the original. Forks typically denote a deviation from the original project that is unlikely

[1] `https://github.com/piotrmurach/github`

to be reconciled. Finally, pull requests facilitate external developers making small changes which tend to be fixes to problems found or desired feature implementation. The owner of the original repository can then decide to integrate the changes made the original repository.

3.2 Storage

As mentioned above the data is stored in a MySQL relational database which leverages Structured Query Language (SQL). There are three databases used for the collection and the analysis. The first stores the raw mined data, whereas the second stores the analyzed data in a more convenient layout to be used later. Finally, the third database stores the same data as the second however uses a different relational database implementation because of some limitations within MySQL. This third database uses PostgreSQL, which has a more advanced set of features than MySQL and is simply a clone of the second database. The specific limitations that were encountered will be discussed more fully later in this section.

The first database, *github_data*, stores the semi-raw data collected from GitHub's API. This database contains 8 tables which store various aspects about the projects considered potentially important for the analysis later on. The tables of primary concern are *repositories*, *commits*, *users*, *files* and *tags* tables. The complete schema is outlined in Figure 3.1. Other aspects are available from the API and if needed the database could be extended to store more elements as necessary. In some cases, data from the API is not available for one reason or another (usually inaccessible files or such) these are simply removed or a note is made of them depending on their importance. For example, non Java source files that are missing are not essential and if inaccessible are ignored. If a Java file is inaccessible, a note is made as this is a greater

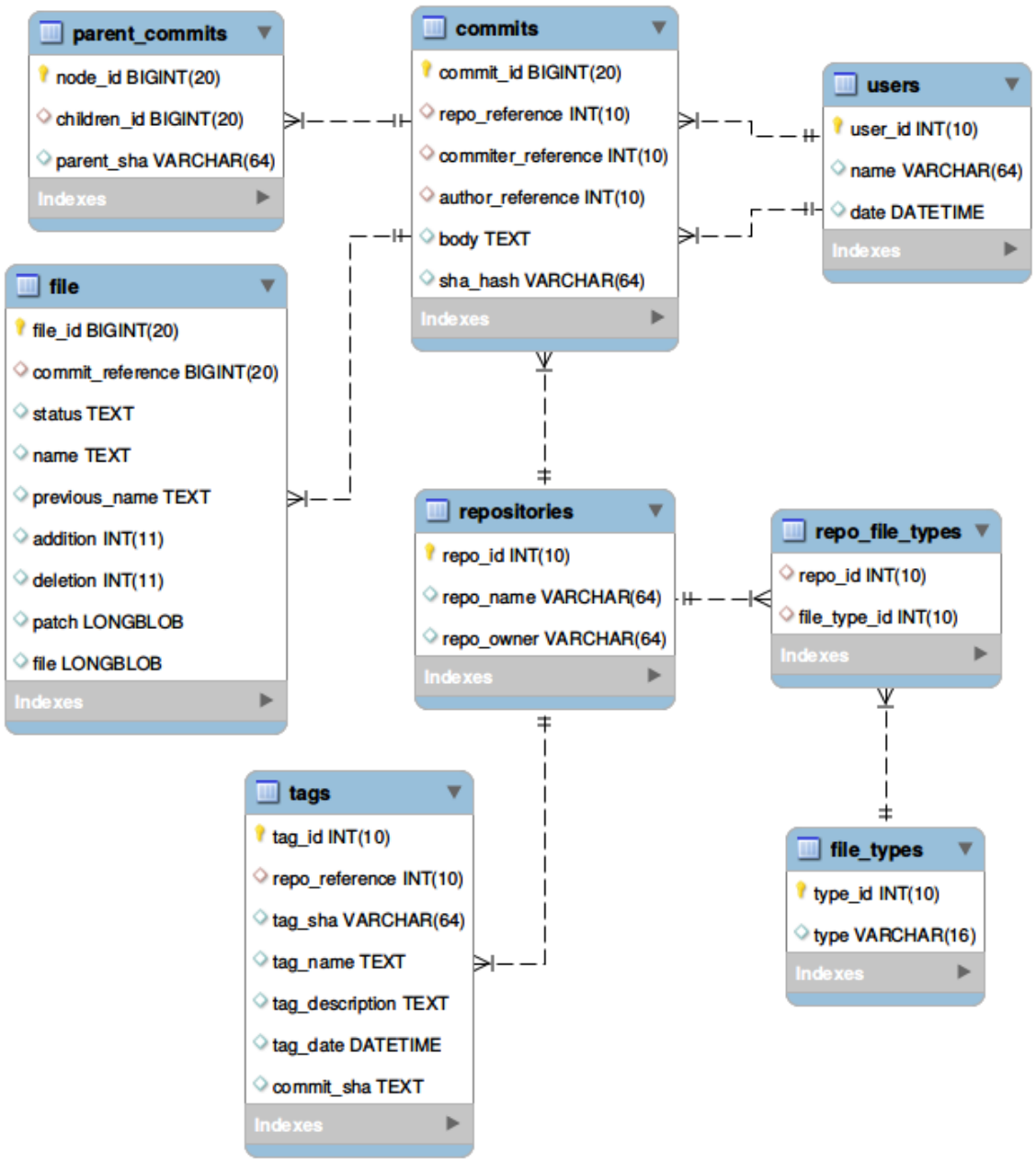


Figure 3.1: GitHub Data Schema

concern. These files can be retrieved if enough information is available (previous version and corresponding patch file). In the case that insufficient information is available, the analysis can still be applied but will likely adversely effect the result.

After storing the data in the *github_data* database, the analysis process is done.

The *parsing* script is run next and discussed further in the section 3.3. The results are then stored in the *project_stats* database that is very similar in layout to the first database except some extra tables have been added and a few data items have been removed. Mostly the storage expansions are designed to hold change information calculated from the analysis of the data. The complete schema is outlined in Figure 3.2.

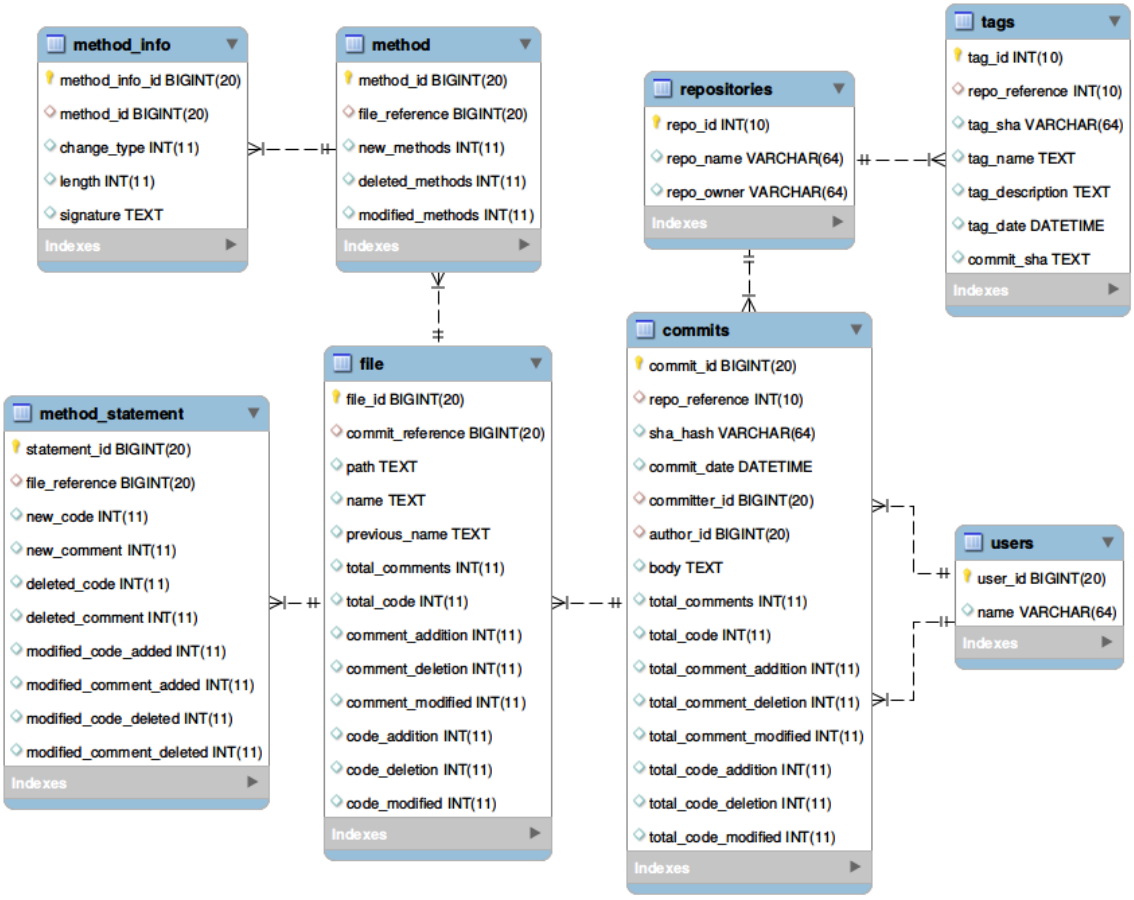


Figure 3.2: Project Stats Schema

The third and final database uses PostgreSQL because of limitations within the MySQL implementation. The calculation of the candidate features, discussed in further detail in section 5.1, required a more versatile partitioning function and the ability to perform multiple inner queries. The first of which is more difficult to

implement and the second is not available at all MySQL. Therefore the data was transferred over to PostgreSQL, using simple program called *pgloader*[2]. Only one difficulty was encountered during the transferring process. One of the tables in the MySQL database was called *user*, however in PostgreSQL, this is a reserved table name and therefore the table cannot be interacted with properly. The work around was to simply rename the table in MySQL prior to transferring to avoid any issues with the database. After transferring the data over to PostgreSQL, the data change predictions are ready to be preformed.

3.3 Parsing

The raw data collected from GitHub is stored and undergoes an analysis to extract more refined details. The process first requires the changes from a commit, the patches, to be merged into their corresponding full file. A patch is simply a stub file which summarizes the changes that occurred within a source file. Once the patch is merged with the raw source file, a full file is formed that contains every change as well as the source code that did not change. Within a patch file and a full source file, three different types of changes are present; additions, deletions and no change. These are represented as a plus sign, minus sign and space respective. An example is outlined of each of these changes in Figures 3.3, 3.4, 3.5, 3.6. The coloring used within these images is purely for visual effect and not present in raw patch files.

The process of reconstructing a full file using a patch file requires modifying the original source file. Since the source file used is the product of the patch file, the patch must be applied in reverse. Therefore lines with a plus sign, additions, are assumed to be within the file and lines with minus signs, deletions, are assumed to not be

[2] `http://pgloader.io/`

```
+    private static <E, T extends Comparable<T> > Integer binarySearch
+        (ArrayList<HashMap<E, T > > patterns, T target, E attribute,
+        int start, int end) {
+
+        if(start > end) return null;
+
+        // One element left lets check that element
+        if(start == end) {
+            if(patterns.get(start).get(attribute).compareTo(target) == 0) {
+                // The last value is the one we are looking for
+                return start;
+            }
+            return null; // The value is not here.
+        }
+
+        int middle = (start + end) / 2;
+        int result = patterns.get(middle).get(attribute).compareTo(target);
+        if (result > 0)
+            return binarySearch(patterns, target, attribute, middle+1, end);
+        else if(result < 0)
+            return binarySearch(patterns, target, attribute, start, middle-1);
+        return middle;
+    }
```

Figure 3.3: Newly added method

present within the file. The lines previous removed from the source file are added back to their original location with a minus sign to preserve the original meaning of the line. The lines that were added to the source file are perpended with a plus sign to identify that the line is an addition.

This full source file is then analyzed to extract each method to identify the type of change as well as other method metrics. The type of change that occurred to a method is identified as one of four possible changes. First, a method may be completely new and is thus classified as a new method shown in Figure 3.3. The second method closely, related to the first would be an entirely removed method that is classified as a deleted method shown in Figure 3.4. The third classification that is more difficult method to identify is a modified method. Simply a modified method

```
-    private static <E, T extends Comparable<T> > Integer binarySearch
-        (ArrayList<HashMap<E, T > > patterns, T target, E attribute,
-        int start, int end) {
-
-        if(start > end) return null;
-
-        if(start == end) {
-            if(patterns.get(start).get(attribute).compareTo(target) == 0) {
-                // Value Found
-                return start;
-            }
-            return null; // The value is not
-        }
-
-        int mid = (start + end) / 2;
-        int result = patterns.get(mid).get(attribute).compareTo(target);
-        if (result > 0)
-            return binarySearch(patterns, target, attribute, mid+1, end);
-        else if(result < 0)
-            return binarySearch(patterns, target, attribute, start, mid-1);
-        return mid;
-    }
```

Figure 3.4: Removed method

is one that contains at least two of following three change types; added, removed or unchanged lines. An example of a method that contains all three change types is shown in Figure 3.5. In the event that a method consists entirely of additions and deletions then the method is classified as both a new method and deleted method. A deleted and added classification is used over a modified classification because if all lines are deleted and re-added then the method is change far more drastic than a simple modification. The final change type is that of no change, where the method does not contain any changes and is shown in Figure 3.6.

For each commit, this information is stored to allow for easier access and save time since the analysis of larger datasets can be time intensive. In order to maintain the integrity of the initial dataset, this information is stored in a new database. There

```
    private static <E, T extends Comparable<T> > Integer binarySearch
        (ArrayList<HashMap<E, T > > patterns, T target, E attribute,
        int start, int end) {

        if(start > end) return null;

        if(start == end) {
            if(patterns.get(start).get(attribute).compareTo(target) == 0) {
                // Value Found
                return start;
            }
            return null; // The value is not
        }

-       int middle = (start + end) / 2;
-       int result = patterns.get(middle).get(attribute).compareTo(target);
+       int mid = (start + end) / 2;
+       int result = patterns.get(mid).get(attribute).compareTo(target);
        if (result > 0)
-           return binarySearch(patterns, target, attribute, middle+1, end);
+           return binarySearch(patterns, target, attribute, mid+1, end);
        else if(result < 0)
-           return binarySearch(patterns, target, attribute, start, middle-1);
-       return middle;
+           return binarySearch(patterns, target, attribute, start, mid-1);
+       return mid;
    }
```

Figure 3.5: Mixed changed method

are several other features available in the data set from the extraction process beyond the ones outlined here in detail. A few of those features include: the commit author, the commit message and the method length per commit. This data is stored in the database to help create the prediction model later on.

```
private static <E, T extends Comparable<T> > Integer binarySearch
    (ArrayList<HashMap<E, T > > patterns, T target, E attribute,
    int start, int end) {

    if(start > end) return null;

    // One element left lets check that element
    if(start == end) {
        if(patterns.get(start).get(attribute).compareTo(target) == 0) {
            // The last value is the one we are looking for
            return start;
        }
        return null; // The value is not here.
    }

    int middle = (start + end) / 2;
    int result = patterns.get(middle).get(attribute).compareTo(target);
    if (result > 0)
        return binarySearch(patterns, target, attribute, middle+1, end);
    else if(result < 0)
        return binarySearch(patterns, target, attribute, start, middle-1);
    return middle;
}
```

Figure 3.6: Unchanged method

3.4 Visualization

3.4.1 Line Change

The key features are extracted from the data set after performing the collection and analysis. Visualizations were used in order to to better understand resulting data. The first visualization simply showed the changes recorded on a per line basis. These changes were divided into several closely related subcategories of additions, deletions and modifications. Additions identify changes that are new and do not have a corresponding set of deleted code. Similarly, deletions refer to changes that remove lines of code without a corresponding set of additions. Finally modifications are a set of changes which contain a set of additions and deletions that are related.

The relationship between two sets of additions and deletions is determined through the Levenshtein Distance (LD) formula. The LD calculation will determine the edit distance between two strings, where edit distance is defined as the number of characters difference between two different strings. For example, the LD for *happy* and *mapper* would be 3, since h would be changed to m, y to e and r would be added at the end. Normalization is used to allow for more general use of LD for comparing strings of different length. To calculate Normalized Levenshtein Distance (NLD) the LD would be divided by the larger of the two strings sizes shown in Equation 3.1.

$$NLD(a_i, d_j) = \frac{LD(a_i, d_j)}{\max(|a_i|, |d_j|)} \tag{3.1}$$

Line modifications are assumed to only take place in a series of line changes that involved both additions and deletions shown in Figure 3.5. In this example, 3 line modifications take place each containing 1 addition and 1 deletion. A line modification can also have an $|a|$ to $|d|$ relationship. That is to say more generally, $|a|$ lines of addition may relate to $|d|$ lines of deletion where both $|a|, |d| > 0$. In order to determine whether two lines are closely related enough, a threshold Δ_m is defined. As outlined in Equation 3.2, when the NLD is below the threshold Δ_m then the two lines are related.

$$m(a_i, d_j) = NLD(a_i, d_j) < \Delta_m \tag{3.2}$$

Normalizing the LD calculation accounts for the differences in line sizes when being compared. With shorter lines, the change of a variable name could change a large portion. Therefore with smaller lines likely modifications result in a dramatically higher distance between lines. Likewise, longer lines can contain more text modifications and still result in a low score because of the length of the line. This resulted

in the creation of the a threshold α to separate small and large line changes. The Equation 3.2 is updated accordingly shown in Equation 3.3.

$$m(a_i, d_j) = \begin{cases} NLD(a_i, d_j) < \Delta_s & \text{if } \max(|a_i|, |d_j|) < \alpha \\ NLD(a_i, d_j) < \Delta_l & \text{otherwise} \end{cases} \tag{3.3}$$

Only lines that are part of the same block of additions and deletions are selected for the similarity check to determine whether they can be classified as a modification. As noted before, line modifications will consist of one to many addition lines mapped to one to many lines of deletion. Therefore a modification is more easily referred to as a modification set. For addition lines that do not meet the threshold of similarity with all deletion line in the change block are classified as additions. Similarly, deletion lines that fail to meet the similarity threshold for all addition lines will be classified as deletions. A block of changes will therefore contain a set of added and deleted lines, some of which may be related.

While creating the parsing method, both code and comments were considered separate entities. However each was analyzed with the same method. Therefore for the visualization below, the changes are separated into source code and comments. The comments were never used towards the prediction method presented later on in chapter 4 and will therefore not be covered as deeply. The comments however are available for use and could be used to extend the approach.

Each of the following visualizations presented have a interactive component which is not available. Their capabilities are summarized to provide context to their use. The line change data is visualized in Figure 3.7. The number of changes lines of source code added, deleted and modifications are all shown aggregated per month. A per commit view is available but is very cluttered because of the excess of data with in a project. The bottom half of the visualization shows the sum of changes

up till the given point. Tags for the project are shown at the bottom of the graph to provide some context of the release cycle. Tags often mark points of significance within the project and therefore can be thought as road signs. The visualization also provides some options to refine or generalize the view. For all of the views, the user is allowed to select the project, package path, and the committers as desired. Specifically for the line level graph, a further option is provided to condense the data based on a monthly, weekly and commit summary. The commit message and a link to the commit on GitHub are provided when viewing either the commit view, method level or statement level. This information allows for a direct link to the project and can be a handy tool for referring back to the software repository.

Each view is also supplemented with a summary of key stats for the project. In Figure 3.8 the project stats are outlined for *acra*. The ratio of comments to code for the history of the project is shown in a pie chart. Several table entries outline several top performers metrics are outlined with the top five for each category. The first set outlines top contributers for source code, which is broken down into the categories: coder, modifier and deleter. Coder, modifier and deleter all map to the top committer to provide additions, modifications and deletions respectively. The number of line added is also outlined as well next to the committer's name in square brackets. The second set of lists outlines the top commenters for additions, modification and deletions. Finally, the top committers and contributors are listed, where the committer is sum of commits contributed to the project, and contributer is the sum of total code and comments contributed.

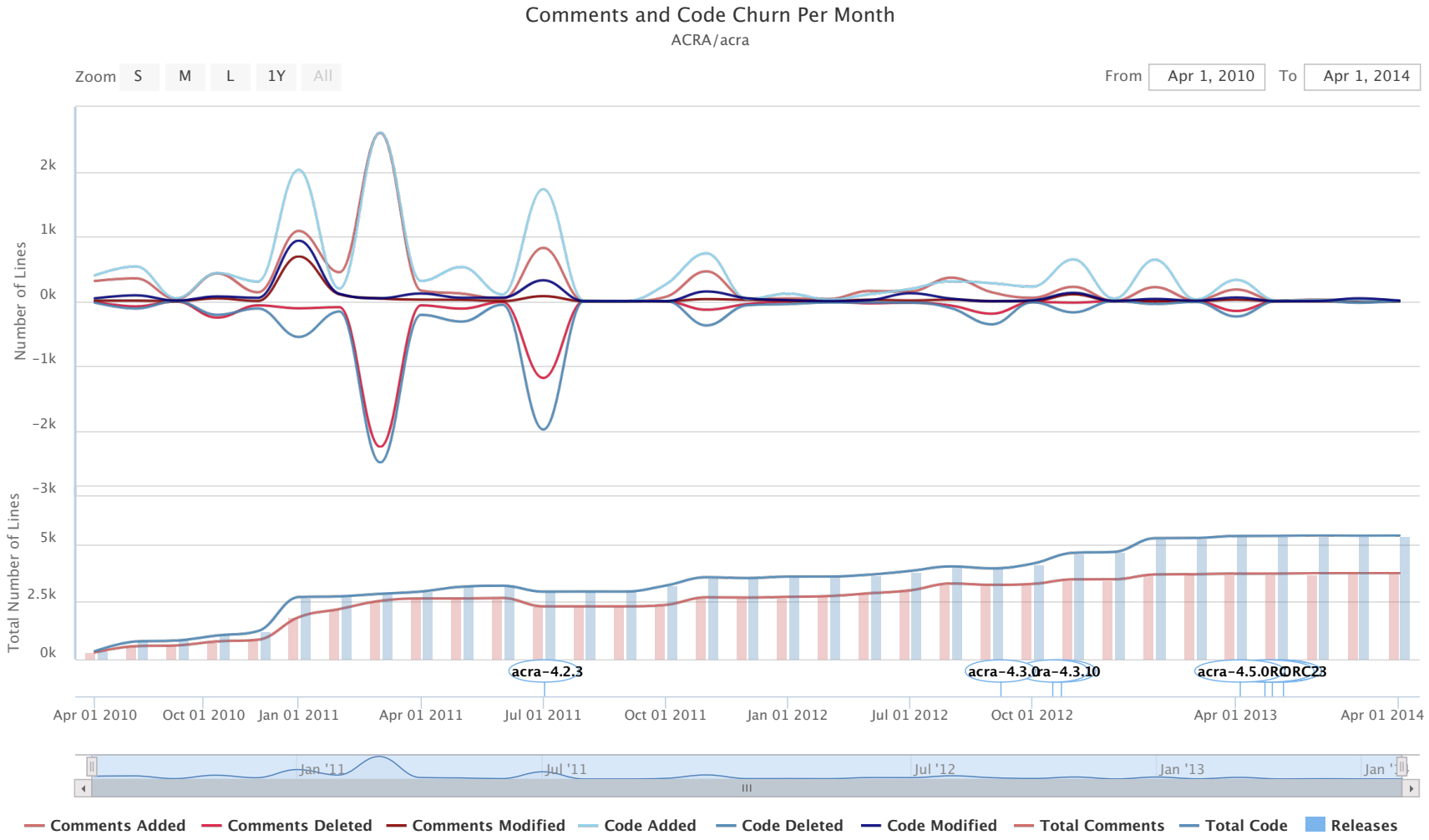
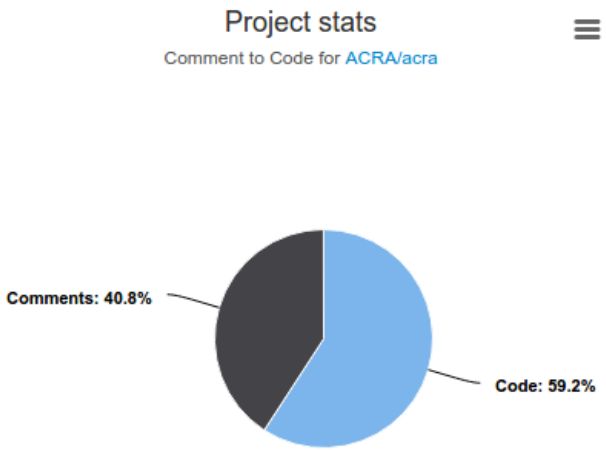


Figure 3.7: Line Change Visualization for acra



Top Committers	Top Contributors
1. KevinGaudin [*573*]	1. KevinGaudin [*8569*]
2. William Ferguson [*133*]	2. William Ferguson [*-720*]
3. pyricau [*48*]	3. william-ferguson-au [*341*]
4. william-ferguson-au [*19*]	4. npombourcq [*-42*]
5. npombourcq [*5*]	5. Richard Meinsen [*22*]

Top Coders	Top Modifiers	Top Deleters
1. KevinGaudin [*8388*]	1. KevinGaudin [*2182*]	1. KevinGaudin [*3126*]
2. pyricau [*2469*]	2. William Ferguson [*332*]	2. pyricau [*2469*]
3. William Ferguson [*1826*]	3. npombourcq [*43*]	3. William Ferguson [*2089*]
4. william-ferguson-au [*358*]	4. william-ferguson-au [*29*]	4. Richard Meinsen [*104*]
5. Richard Meinsen [*126*]	5. meddins86 [*13*]	5. william-ferguson-au [*40*]
Top Commenters	**Top Comment Modifiers**	**Top Comment Deleters**
1. KevinGaudin [*5502*]	1. KevinGaudin [*1191*]	1. pyricau [*2219*]
2. pyricau [*2219*]	2. William Ferguson [*80*]	2. KevinGaudin [*1461*]
3. William Ferguson [*840*]	3. meddins86 [*3*]	3. William Ferguson [*1196*]
4. Richard Meinsen [*96*]	4. william-ferguson-au [*2*]	4. Richard Meinsen [*96*]
5. william-ferguson-au [*34*]	5. fernandohur [*2*]	5. william-ferguson-au [*8*]

Figure 3.8: Project Summary Statistics for acra

3.4.2 Method Change

The visualization of line changes was very noisy and proved difficult to use. Instead of viewing every line of change separately, the changes are grouped together based on the method from which they originate from. Similar classifications are used for method changes however their definitions vary slightly and are outlined in more detail in section 3.3. There are three types of method level changes that can occur. Firstly, a method is classified as newly added when that method had not existed in the previous version, consisting only of additions. Secondly, a deleted method implies that the method is completely removed from the current version, consisting only of deletions. Thirdly, a method is classified as modified if it contains two or more types of changes, either additions, deletions or no change.

The method level change visualization, shown in Figure 3.9, presents the amount of method changes that occurs in the project development over time for acra. The low level changes details are ignored in this view, instead the focus is placed on that of the three types of changes. The visualization for the method level uses a bar graph since it provided a more clear picture of the relationship between commits. Compared to the first visualization which implied that a relationship between different commits of the same type of changes. The contrast in magnitude between each type of change and each commit is also more clear and defines the visualization. The amount of change occurring over time is clearly visible and the amount of data available is not as overwhelming as the line change visualization.

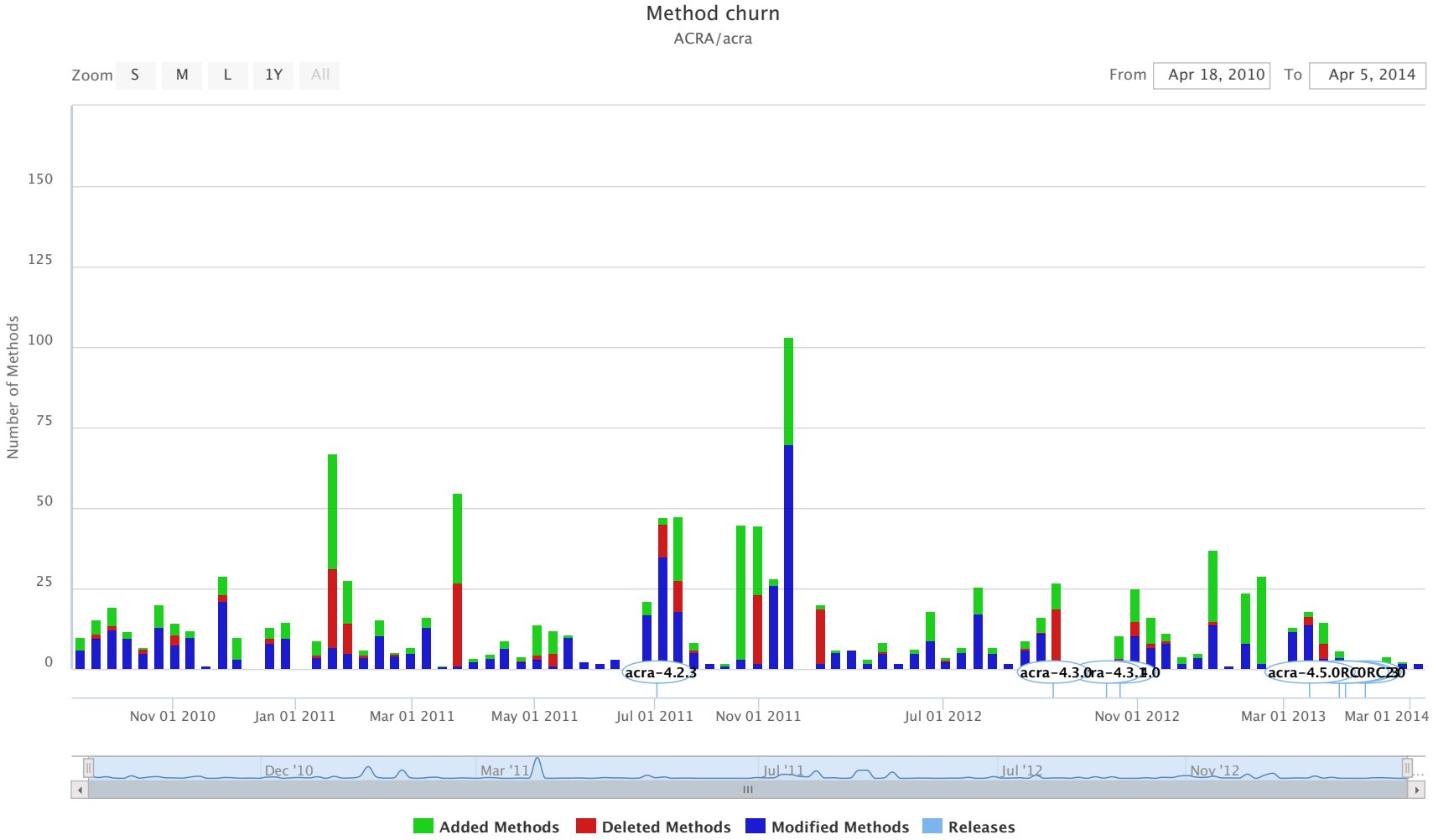


Figure 3.9: Method Change Visualization for acra

3.4.3 Method Statement Change

The method statement level visualization is a more granular view of the method level view. By building on top of the method level view, the method statement level view provides more details similar to that of the line level view. The classifications are kept from the method level view for changes but are broken down into line changes made to comments and code. Therefore both methods consist of two parts, the comments and the code. So the previous classification added, deleted and modified are divided into new categories. Added and deleted methods are divided into two new categories each; added code, added comment, deleted code and deleted comment respectively. In Figure 3.10, the added and removed method data is shown. The modifications are not shown in the visualization to reduce clutter in the view.

In Figure 3.11 the modifications for method statement changes are shown for acra. Modifications are divided into four categories instead of two. The first two categories relate to modification of comments, and the second two relate to modifications of code. The comments changes are classified as either modified added or deleted comments. Likewise for source code modifications, the classification is either modified added or deleted code. Each line classified under modification is based on the change type of the method. Therefore if a line of source code is part of a method that is modified then it will fall into one of the four modification classifications. For example in Figure 3.5, there would be 5 lines classified as a modified deleted code, 5 lines of modified added code and 0 lines of modified added or deleted comments.

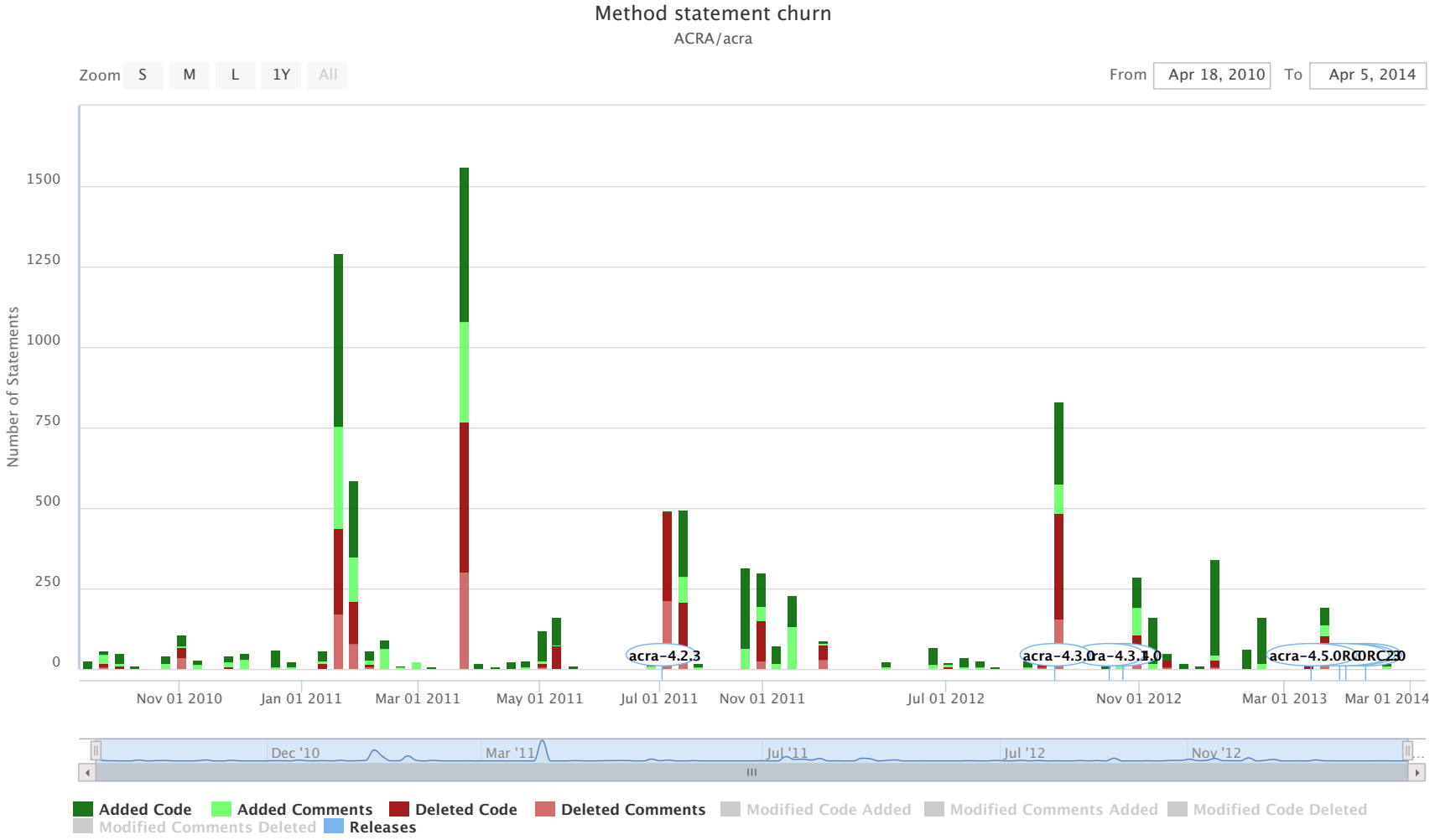


Figure 3.10: Method Statement Added & Deleted Visualization for acra

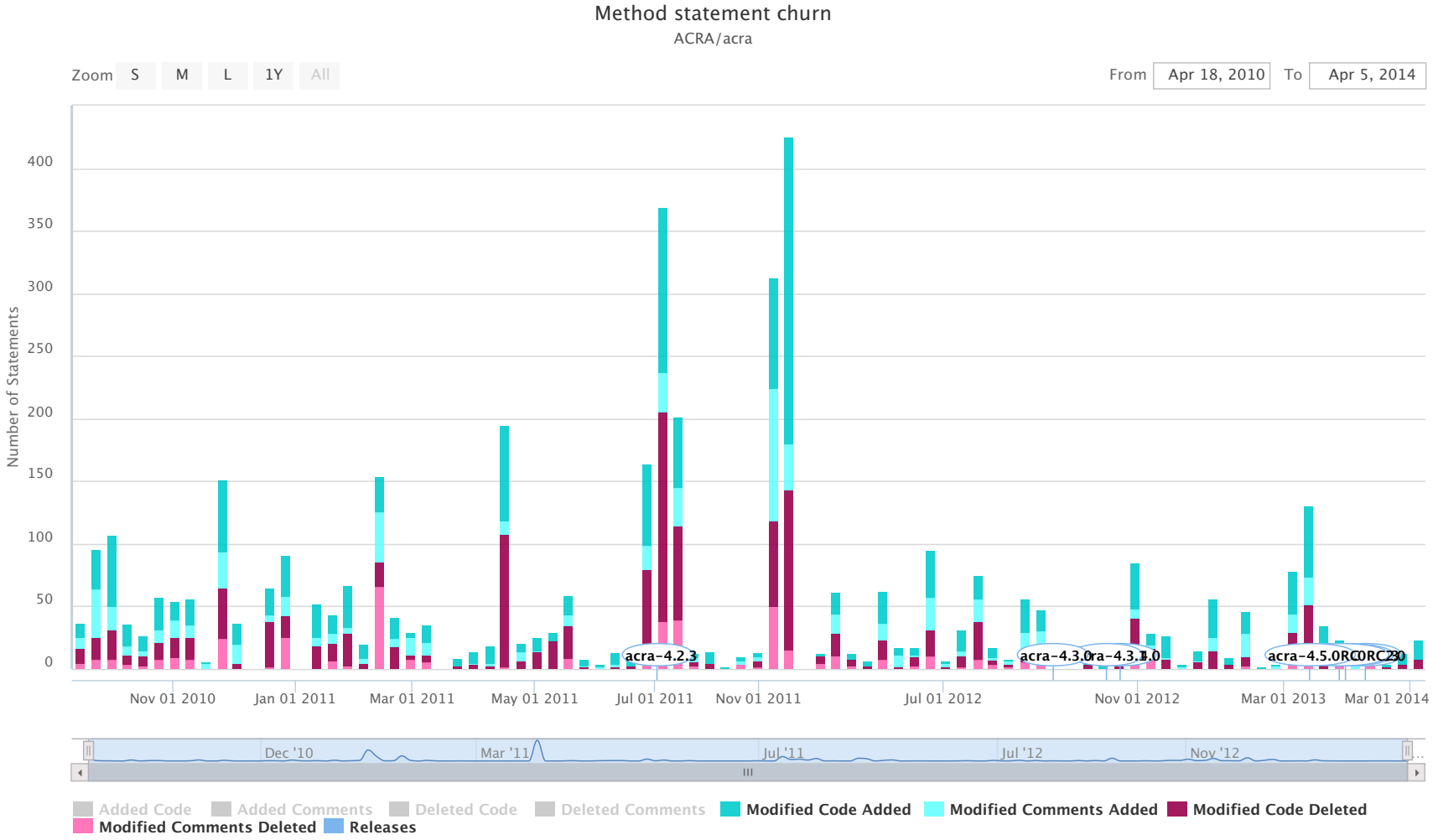


Figure 3.11: Method Statement Modification Visualization for acra

Chapter 4

Prediction with Commit Data

The visualization for the data collected from the projects helps provide several insights into data set which can be used to help with the creation of the prediction scheme. With the data visualized, a more general look of the data collected is available. While creating the method for predicting change within the project, the visualizations provide a helpful resource. The visualization can help identify relationships between variables and general trends. The actual data used for training the prediction model is outlined in section 4.1. After that the prediction model is detailed in section 4.2.

The data presented in the visualization is used towards creating an approach to predict whether a method will change within the next five commits. The machine learning algorithms used in the approach are SVM and RF. Of course the performance of a prediction method will be influenced by several factors including; the size of the sample, the features used for training and balancing of the data set.

4.1 Prediction Data

The data used to predict changes within a project is originally from the visualizations. For more information about the specific information collected see: section 3.1, section 3.2 and section 3.3. The commit data collected from the target OSS project is used to make predictions. The goal is to predict whether a method within a project will change within the next five commits. As outlined in section 3.3, the different types of changes can be either additions, deletions, modifications or no change at all.

The machine learning model requires samples from the data set to train from which allows for predictions of new elements. The training samples taken must also be categorized based on the desired outcome of the machine learning algorithm. This requirement provides some restrictions on which values can be included in the sampling. Since the categorization is whether a method will change, all methods sampled need to be able to collect data the next five commits following the current commit. Therefore methods that are within the last five commits of the training sample window are not included in the training set for the data model.

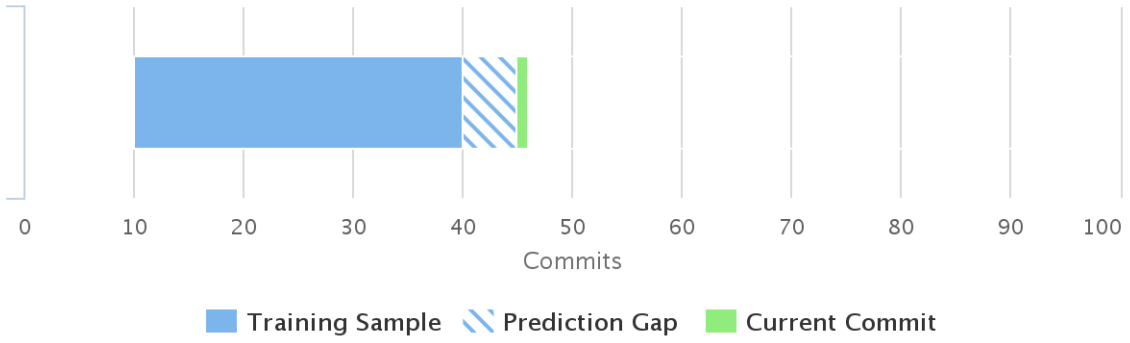


Figure 4.1: Training Sampling Layout

As noted above, one of the key factors in the performance of the prediction approach is the size of the sample set. The sample set size is restricted by a variable

value SWR which controls the number of commits considered to sample from. The data will only be sampled within a limited range of commits as outlined in Figure 4.1. In this case, the SWR is 30 and the prediction gap is 5. The prediction gap takes into account the categorization restriction, preventing sampling from the last five commits.

Another consideration when sampling the data is the distribution of the categories. Most of the time, a data set will contain more samples in one category than the other. For example, a sample may contain 80% methods with no change within the next five commits and 20% methods with change. Ideally the number of methods with changes and without changes in the next five commits would be close to 50%. When the distribution of samples for each category is very close the model will often perform better. However in cases where the data is highly skewed to one classification over the other, the model will often predict the larger classification for input values. OS will re-samples from the smaller classification to reduce the size difference with the larger classification. Alternatively, undersampling the data set will remove samples from the larger classification to reduce the difference in size with the smaller classification.

Undersampling is applied to a data set by measuring the number of samples in each category. The larger of the two is reduced by discarding samples at random until the data set is the same size as the smaller data set. This will reduce the number of samples used to train the model and may reduce the performance of the model based on the decreased number of samples. In cases where there are a limited number of samples for the smaller category, undersampling may not be ideal. OS alternatively increases the number of samples by calculating the number of samples in each category and expanding the smaller category by re-sampling values from the data set until both categories are equal in size. Both approaches can be used together so that the smaller category is expanded to at most twice its original size. If

the initially smaller category is still smaller, the larger category set will be reduced to the size of the smaller category. When selecting values for re-sampling or removal, a randomized selection process is used to ensure the distribution of the data is preserved. For example, if category a, with $|a| = 100$ and category b with $|b| = 1000$. OS will be applied to category a since $|a| < |b|$. Therefore a will apply random re-sampled until $|a_n| = |a| \times 2$ or $|a_n| = |b|$. Once one of the conditions is met, OS is complete. Next undersampling is applied to the larger category b, where samples are randomly removed from b until $|b_n| = |a_n|$.

Once the categories are balanced then the model can use the data. However with large sample sizes, a reduction of the sample set may be necessary. The variable $sample_r$ is the percentage of the number of samples taken from the range. Instead of picking an arbitrary number of samples, a ratio was used to scale based on the number of available samples. When sampling, if the ratio is at 50% then only half of the values retrieved will be used to train or test. For some of the larger data sets sampling 100% of the data from the range would take a lot longer. Therefore sampling a percentage of the data set is commonly used to decrease the training time. In order to provide a more stable model, a random sample of the sample range is used so that each data entry in the sample has the same chance to be within the training or test data set. Using the example from above, a and b have been oversampled and undersampled such that their new sizes are represented by $|a_n|$ and $|b_n|$ respectively. Given that a sample ratio of 50% is used then both sets a and b would be reduced by the ratio by randomly sampling from each set to create new sets. The size of each set would be $|a_n| \times r$ where r is the ratio value.

The Table 4.1 outlines each of the considered features used for training the prediction model. An example of each feature is provided to further illustrate the feature. As stated in the previous subsection 2.2.1, the values need to be processed into a

Feature	Description	Data	Example Vector
Com	The individual who committed the change	bob	5
Sig	The method signature related to the change details	void getValue()	46
Name	The name of the file	Main.java	3
Δ_i	Whether the method changed or not in the current commit	3	1
m_+	Whether the is newly added	3	0
m_-	Whether the method was deleted	3	0
m_c	Whether the is a modification	3	1
m_x	Whether the received no change	3	0
Δ_{i-j}	Whether the method changed in a previous commit	0	0
$type_{\Delta_{i-j}}$	Type of method changed in a previous commit	2	2
f_Δ	The frequency that the method is changed within the SWR	0.0464	0.0464
sf_Δ	The frequency that the method is changed within the last 10 commits.	0.1	0.1
t_Δ	The time between the current commit c_i and the previous commit c_{i-1}	2148	2148
$t_{\Delta_{i-j}}$	The time difference between a sequence of two previous commits	453	453
Length	The length of the method in this commit	10	10
$change_{t-1}$	Whether a change has occurred in the previous 5 commits	$\{3, 0, 0, 3, 0\}$	1
$change_t$	Identifies whether a change occurred within the next 5 commits for the given method	0	0

Table 4.1: Candidate features for SVM model

usable format for SVM or RF. First the data is extracted from the database as *raw* values as shown in the ***Data*** column. Text values are mapped to a integer value. For example the *Name* value, "Main.java" will be mapped to the value 3. The reason the value is 3 is because 2 other methods have already been mapped and therefore method name is mapped to the next available mapping. Similarly both *Com* and *Sig* will be mapped from their respective values "void getValue()" and "bob" to 46 and 5. Numerical values are converted by casting the value to a floating point value if the value is not that type already. For spacing reasons, all the values in the table that have no decimal value are shown without a ".0" following.

Several experiments were conducted to investigate the value of each of the candidate features. The candidate features were narrowed down from this initial list into a smaller set of training features. These features sets were constructed to determine potential ideal feature sets. The complete list of features used are shown in Table 4.2. This list is not ordered by execution order, rather the first five are kept consistent with the numbering used in the Some of the feature sets are not full explained in the table for spacing reasons. Feature sets 16, 17, 18 and 27 all use the last five previous changes or durations. Each are marked similarly to those that only use the most recent previous change or duration except for a footnote marker. Likewise feature set 29 is also different since the previous change used is only the change made five commits ago. Each of these feature sets are tested on the repository acra and the results are shown in Figure 4.2.

Clearly feature sets 1, 2, 4 and 5 all perform better overall with all three measures performing very high together. For a number of cases the feature set performed well for one or two of the three performance measure but performed poorly in the rest. Therefore the first five were selected for experimentation of the effect of feature sets on various projects using SVM and RF. Feature set 3 did not perform as well as the

other selected feature sets but some preliminary experimentation had shown other projects vastly differently than acra. The goal was to select a feature set that did not perform as well for acra but still had fairly high performance to potentially perform better for other projects.

Feature Sets	Com	Sig	Name	Δ_i	m_+	m_-	m_c	m_x	Δ_{i-j}	$type_{\Delta_{i-j}}$	f_Δ	sf_Δ	t_Δ	$t_{\Delta_{i-j}}$	Length	$change_{t-1}$
1	•	•	•								•	•			•	•
2	•	•	•								•		•		•	•
3	•	•	•								•		•			•
4		•	•								•		•			•
5	•	•	•								•					•
6	•	•	•	•							•					•
7	•	•	•		•						•					•
8	•	•	•			•					•					•
9	•	•	•				•				•					•
10	•	•	•					•			•					•
11	•	•	•						•		•		•			•
12	•	•	•							•	•		•			•
13	•	•	•							•	•					•
14	•	•	•						•		•					•
15*	•	•	•								•			•		•
16†	•	•	•						•		•					•
17‡	•	•	•							•	•					•
18*‡	•	•	•							•	•		•			•
19	•	•	•		•	•	•				•					•
20	•	•	•			•	•				•					•
21	•	•	•			•		•			•					•
22		•	•								•					•
23		•	•								•					
24		•	•						•		•					
25		•	•	•							•					
26		•	•	•							•		•			
27†		•	•	•					•		•					
28		•	•	•					•		•					
29§		•	•	•					•		•					

Table 4.2: Training Features

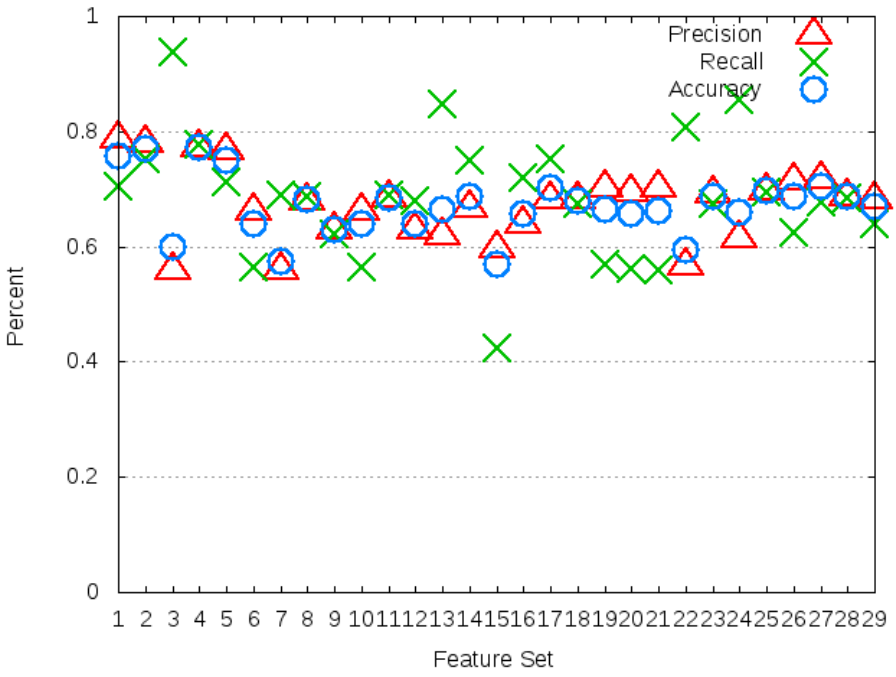


Figure 4.2: Feature Sets Analysis using RF

Another small change made to the data to create a vector for the prediction model was to convert the change type into a change indicator vector using Equation 4.1. The vector is converted into a single value which indicates whether a change has occurred in the previous five commits. This process is done through calculating the sum of the change vector using Equation 4.2. Finally, the change indicator, $changei - 1$, is identified using Equation 4.3.

$$C = \begin{cases} 1 & \text{if} change > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

$$reduce = \sum_{i=t-5}^{t} c_i \tag{4.2}$$

---

*$t_{\Delta_{i-j}}$ denotes the use of the 5 previous change duration for this feature set.
†Δ_{i-j} denotes the use of the 5 previous changes for this feature set.
‡$type_{\Delta_{i-j}}$ denotes the use of the 5 previous change types for this feature set.
§Δ_{i-j} denotes whether change occurs at the fifth previous commit for this feature set.

$$P = \begin{cases} 1 & \text{if} reduce > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

f_Δ is calculated by taking the number commits in which involve changes to the current method (c_i) within the SWR divided by the current number of commits (c_{cur}) since the start of the SWR. This is formalized in Equation 4.4. A frequency of change is available for the duration of the SWR.

$$f_\Delta = \frac{|c_i|}{|c_{cur}|} \tag{4.4}$$

sf_Δ is calculated by reducing the range sampled to s. Then counting the number of times the method changes within the last s commits and dividing it by s. The size of the short frequency can be any value that is less than the size of the SWR. For use in the rest of the paper $s = 10$ which means that the sf_Δ is for the last 10 commits.

t_Δ is the difference between the current commit time ($t(c_i)$) and the previous commit time ($t(c_{i-1})$) calculated in Equation 4.5. Both time values are provided as time stamps and the result is calculated in seconds. Only the difference in time between the current commit and the previous one is calculated therefore, in Equation 4.5, i denotes the current commit.

$$\Delta t_i = t(c_i) - t(c_{i-1}), i > 1 \tag{4.5}$$

4.2 Prediction Method

For this approach a machine learning algorithm is used to create a prediction model. The data used to train the model is collected as shown in section 4.1. The machine learning algorithms that can be used are either SVM or RF. Each of these method

are widely used for data mining techniques and are easy to use. Figure 1.1 outlines the overall structure of the approach for how changes are predicted.

The SVM model was created through the use of a libsvm[1] binding for Ruby, rb-libsvm[2]. This library was a good fit since the data was collected using a Ruby script. For RF python library scikit-learn[3] was used. The reason for switching from Ruby to Python was a lack of a mature library for RF in Ruby.

Use of the approach requires a few steps in total before predictions can begin. Firstly, the data related to the project must be collected from GitHub. Once the data is collected, the prediction model can be created by providing training set built from sampling the data set. After training the model, predictions can be made using the model on new data. The next chapter analysis the approach by training and testing the model using subsets of the project data sets.

[1] `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`
[2] `https://github.com/febeling/rb-libsvm`
[3] `http://scikit-learn.org/stable/`

Chapter 5

Experiments

Experiments were conducted in order to provide some validation for the proposed approach. The goal of these experiments is to test the whether the historical commit data from an OSS project can be used towards predicting future changes within the project. These experiments are based on the approach outlined in the previous chapter, chapter 4. The experiment was conducted through measuring the performance of each core factors varied in isolation. Specifically the SWR, the feature set and categorization balancing using OS. These three factors are explored for both machine learning algorithms.

5.1 Experimental Project Data

The complete list of projects that were experimented on are found in Table 5.1. Data from each project was collected from the creation date for the project till the date the data was collected on. The number of commits excludes any commit that lacked a change to a file containing Java code. Since the primary interest was to parse Java code, files containing Java code were used while all other files are ignored.

These measures provide a more accurate description of the project in terms of the analysis and predictions made on it. Secondly, the number of developers does not map effectively to what git uses as committers and authors. Instead, the number of developers includes all individuals (removing duplicates) who committed or authored commits to the current project.

Owner	Project	Start Date	End Date	# of Commits	# of Developers
ACRA	acra	2010-04-18	2015-06-05	404	32
arquillian	arquillian-core	2009-11-13	2016-03-16	473	49
google	blockly-android	2015-07-23	2016-06-23	691	8
openzipkin	brave	2013-04-07	2016-06-21	337	32
gabrielemariotti	cardslib	2013-09-20	2015-05-12	327	13
square	dagger	2012-06-25	2016-01-30	496	38
deeplearning4j	deeplearning4j	2013-11-27	2016-02-13	3523	61
facebook	fresco	2015-03-26	2015-10-30	313	45
Netflix	governator	2012-03-18	2016-06-23	621	31
greenrobot	greenDAO	2011-07-28	2016-05-23	415	4
kevinsawicki	http-request	2011-10-21	2015-01-21	273	14
koush	ion	2013-05-22	2016-06-14	520	29
skylot	jadx	2013-03-18	2016-03-27	480	11
mapstruct	mapstruct	2012-05-28	2016-06-15	604	22
Atmosphere	nettosphere	2012-02-09	2016-04-11	336	12
johncarl81	parceler	2013-07-03	2016-06-22	228	12
orfjackal	retrolambda	2013-07-20	2016-04-30	275	11
amlcurran	ShowcaseView	2012-08-14	2016-05-30	332	39
haifengl	smile	2014-11-20	2016-06-24	237	14
perwendel	spark	2011-05-05	2016-06-19	551	86
apache	storm	2011-09-16	2015-12-28	2445	260
prestodb	tempto	2015-03-06	2016-06-20	298	19
gridgain	yardstick	2014-04-11	2015-10-12	213	12

Table 5.1: Experiment projects

Given the large number of project experimented on a categorization system was necessary to allow for grouping of projects. Four project measures were selected for comparing the projects and are outlined in Table 5.2. The measures are: project

project length	project size	# devs	commit rate
short ($t < 1$)	small ($m < 2000$)	small ($d < 30$)	low rate ($r < 100$)
medium ($1 \leq t < 3$)	medium ($2000 \leq m < 10000$)	medium ($30 \leq d < 100$)	medium ($100 \leq r < 300$)
long ($t \geq 3$)	large ($m \geq 10000$)	large ($d \geq 100$)	high rate ($300 \leq r < 600$)
			very high rate ($r \geq 600$)

Table 5.2: Experiment project summary

length in years, project size in number of methods, number of developers and the rate of commits made in commits per year. The project length represents the number of years the project has been under development for. The size of the project is measured in the number of method signatures within the project since created. The number of developers is tallied from the beginning of the project for this measure. Finally, the rate of commits is the number of commits contributed to the project per year. A yearly rate of commits was sufficient since the majority of the projects had more than one year of development history.

Using the classifications outline in Table 5.2, the projects are grouped and organized with similar projects. In Table 5.3 the projects are sorted by their classification and have dividing lines around similar projects. For example four projects; http-request, nettosphere, parceler and retrolambda are all classified in the same group. Some projects like ion or storm are not grouped in with another project and thus are in a group of their own.

Each of the projects are selected from GitHub using the list of Java projects with a large amount of development. OSS projects were targeted to simplify any usage concerns. Specifically, OSS projects are open and freely available immediately and can be discussed without restriction. Therefore in order to be selected the program had to clearly use an OSS license. Secondly, the project also needed to have at least a 6 months worth of development and at least 300 commits to provide a large enough

name	project length	project size	# devs	commit rate
yardstick	medium	small	small	medium
tempto	medium	medium	small	medium
blockly-android	medium	medium	small	high
fresco	medium	medium	medium	high
http-request	long	small	small	low
nettosphere	long	small	small	low
parceler	long	small	small	low
retrolambda	long	small	small	low
ion	long	small	small	medium
acra	long	small	medium	low
dagger	long	small	medium	low
ShowcaseView	long	small	medium	low
greenDAO	long	medium	small	low
smile	long	medium	small	low
cardslib	long	medium	small	medium
jadx	long	medium	small	medium
mapstruct	long	medium	small	medium
arquillian-core	long	medium	medium	low
brave	long	medium	medium	low
spark	long	medium	medium	low
governator	long	medium	medium	medium
deeplearning4j	long	large	medium	very high
storm	long	large	large	high

Table 5.3: Experiment project summary

Project	# of Methods	# of Methods Changes	Avg # of Commits / Year	Avg # of Methods Change / Commit
acra	1309	3605	67.33	9.51
arquillian-core	5563	6657	59.13	15.2
blockly-android	3608	9679	345.5	14.82
brave	4204	7823	84.25	26.98
cardslib	3940	5122	109.0	16.68
dagger	1827	6314	99.2	13.7
deeplearning4j	29896	82198	880.75	24.33
fresco	3463	4139	313.0	14.73
governator	4229	10946	124.2	19.04
greenDAO	4089	8625	69.17	21.84
http-request	726	1740	54.6	6.72
ion	1678	4347	130.0	8.82
jadx	6012	9322	120.0	19.63
mapstruct	7885	10185	120.8	19.04
nettosphere	1112	2857	67.2	9.01
parceler	1619	3076	57.0	14.72
retrolambda	1111	2588	68.75	9.95
ShowcaseView	927	2672	66.4	8.62
smile	3885	3879	79.0	18.47
spark	3117	9154	91.83	18.27
storm	14599	50037	489.0	24.03
tempto	2422	3386	149.0	11.96
yardstick	512	1216	106.5	6.37

Table 5.4: Project Change Statistics

dataset to analyze. An effort was also made to pick projects of different sizes to provide better tests of various conditions.

In order to get a more detailed understand of the selected projects, numerous measures were taken. These measures also allow for each projects to be compared to each other in terms of the development of each of the projects. For example the size of the project is represented through several measures including: number of commits, methods and developers. Several averages are calculated to help establish how the development occurred within a project during the development. A few examples of

average measurements are the number of commits per year, changes per method.

1. **acra**[1] is an Android bug logging tool used with Android applications to capture information related to bugs or crashes. The information is sent to the developers to help them address the issues that their clients encounter while using there application.

2. **arquillian-core**[2] is a platform for creating automated integration, functional and acceptance tests for Java middleware products.

3. **blockly-android**[3] provides a native implementation of the blockly library for drag and drop development on Android.

4. **brave**[4] provides a Java distributed tracing tool for troubleshooting latency problems and is compatible with Zipkin.

5. **cardslib**[5] is an Android library for creating UI Cards in an Android application.

6. **dagger**[6] from square is a Java application used to satisfy dependencies for classes to replace the factory model of development.

7. **deeplearning4j**[7] is a distributed neural network library that integrates Hadoop and Spark. This application is the largest of the all the projects and provides a large wealth of data to analyze.

8. **fresco**[8] from facebook is the smallest project with the shortest development

[1] `https://github.com/ACRA/acra`
[2] `https://github.com/arquillian/arquillian-core`
[3] `https://github.com/google/blockly-android`
[4] `https://github.com/openzipkin/brave`
[5] `https://github.com/gabrielemariotti/cardslib`
[6] `https://github.com/square/dagger`
[7] `https://github.com/deeplearning4j/deeplearning4j`
[8] `https://github.com/facebook/fresco`

period. This project provides a library for using images on Android to attempt to solve limited memory issues with mobile devices.

9. **governator**[9] is a library of extensions and utilities that enhances Google's Guice to provide injector life-cycle and object life-cycle.

10. **greenDAO**[10] provides an Android based light and fast object relational mapping to SQLite database entries.

11. **http-request**[11] is a library accessing the *httpURLConnection* to make requests and then access the response.

12. **ion**[12] provides asynchronous networking and image loading for Android.

13. **jadx**[13] is a Java decompiler for Android Dex and Apk files.

14. **mapstruct**[14] is an annotation processor for generating type-safe bean mapping classes.

15. **nettosphere**[15] provides a WebSocket/HTTP server based on Atmosphere and Netty Framework.

16. **parceler**[16] is a library for creating serialize code.

17. **retrolambda**[17] provides a backport for lambda expressions implemented in Java 8 to Java 7, 6 and 5.

[9] `https://github.com/Netflix/governator`
[10] `https://github.com/greenrobot/greenDAO`
[11] `https://github.com/kevinsawicki/http-request`
[12] `https://github.com/koush/ion`
[13] `https://github.com/skylot/jadx`
[14] `https://github.com/mapstruct/mapstruct`
[15] `https://github.com/Atmosphere/nettosphere`
[16] `https://github.com/johncarl81/parceler`
[17] `https://github.com/orfjackal/retrolambda`

18. **ShowcaseView**[18] is a library for Android that can highlight and showcase components within the UI of a application.

19. **smile**[19] stands for Statistical Machine Intelligence and Learning Engine and is a machine learning library for Java.

20. **spark**[20] a tiny web framework for Java 8.

21. **storm**[21] from apache real time computational system for continuous streams of data. This project is one of the larger projects and has a large development community.

22. **tempto**[22] A testing framework for SQL databases running on Hadoop.

23. **yardstick**[23] is a framework for creating benchmarks specifically for clustered or distributed systems.

Several average measures were also taken which detail the amount of change that occurs within the project. The average number of commits per project coupled with the average number of changes per commit clearly indicates the amount of changes that are occurring with in the project. The rate at which methods are change provides good insight into the growth of a project. While some changes may involve the addition of new methods, others may include the removal of methods or the modification of methods. The other measures relating to the amount of change occurring with a project on average are the number of methods changed per year and the number of

[18] `https://github.com/amlcurran/ShowcaseView`
[19] `https://github.com/haifengl/smile`
[20] `https://github.com/perwendel/spark`
[21] `https://github.com/apache/storm`
[22] `https://github.com/prestodb/tempto`
[23] `https://github.com/gridgain/yardstick`

Project	Avg # of Methods Change / Year	Avg # of Changes / Method	Avg # of Commits / Developer	Max Commits / Year	Min Commits / Year
acra	600.83	4.52	13.93	119	33
arquillian-core	832.13	2.03	36.38	175	6
blockly-android	4839.5	4.68	98.71	690	1
brave	1955.75	4.24	14.65	108	56
cardslib	1707.33	3.28	46.71	223	3
dagger	1578.5	5.64	16.0	236	4
deeplearning4j	20549.5	5.69	65.24	2018	65
fresco	4139.0	1.49	156.5	313	313
governator	2189.2	4.11	24.84	159	75
greenDAO	1437.5	3.94	138.33	137	5
http-request	348.0	2.56	39.0	108	5
ion	1086.75	3.31	40.0	253	7
jadx	2330.5	2.41	43.64	208	11
mapstruct	2037.0	2.04	54.91	288	7
nettosphere	571.4	4.37	37.33	118	5
parceler	769.0	2.43	45.6	76	41
retrolambda	647.0	3.06	25.0	133	24
ShowcaseView	534.4	5.9	10.71	141	6
smile	1293.0	1.86	19.75	121	6
spark	1525.67	3.92	7.25	171	22
storm	10007.4	5.93	15.47	948	118
tempto	1693.0	1.88	16.56	253	45
yardstick	608.0	3.65	23.67	208	5

Table 5.5: Project Change Statistics 2

Project	Max # of Methods Changed / Year	Min # of Methods Changed / Year	Max # of Change / Method	Max # of Commits / Developer	Min # of Commits / Developer
acra	1503	183	52	229	1
arquillian-core	3421	55	110	420	1
blockly-android	9543	136	90	538	1
brave	3300	1038	83	225	1
cardslib	3340	19	95	285	1
dagger	3374	171	65	157	1
deeplearning4j	35869	4377	345	1987	1
fresco	4139	4139	33	269	44
governator	3324	1066	263	316	1
greenDAO	2971	34	63	367	1
http-request	752	14	50	267	1
ion	2315	24	161	492	1
jadx	3915	248	197	436	1
mapstruct	4462	201	81	334	1
nettosphere	1074	10	46	322	1
parceler	1151	516	31	217	1
retrolambda	1501	212	83	237	1
ShowcaseView	1156	74	70	215	1
smile	1918	872	24	155	1
spark	2818	277	72	277	1
storm	26526	2152	314	622	1
tempto	3073	313	44	66	1
yardstick	1163	53	62	137	1

Table 5.6: Project Change Statistics 3

changes per method. Each of these further outline how the changes are being made to the project on average.

A few of the measures are related to the number of developers. These while provided are not the primary focus. The information provided by tracking developer interactions with each other or the repository could be integrated into future work.

While the purposed method was being developed ACRA's acra project was primarily used for exploring and initial testing of the approach. After experimenting on

acra a few of the potential candidate feature sets were distinguished based on their superior performance. Experiments were then run on other projects using the feature sets that performed better.

5.2 Experimental Setup

The experiment is setup to have a set of parameters that can be set. These parameters will remain constant to observer the difference that the independent variable will have on the dependent variables; precision, recall and accuracy. Each experiments will use one of the parameters as the independent variable. An experiment consists of a set of trails where the independent variable is modified to measure the resulting dependent variables. The results of a single trial or of the set of trails for a project will be referred to as the performance of the approach. In order to reduce specific project confounding factors numerous projects were tested on. This will be discussed further in the discussions section. The experiments section only includes a small sample of the projects that were experimented on. The projects shown are ones that exhibited interesting patterns or results. The complete set of performance figures for every project are shown in Appendix A.

This thesis works to determine whether SVM or RF can be used to effectively predict changes that will occur within the project. To potentially provide an answer to this question the factors that are used for the prediction method are studied. The experiments attempt to determine what impact the different factors will have on the purposed methods. These factors include:

1. The SWR which is the size of range which the samples are taken from.

2. The set of features used to train the machine learning model.

3. The distribution of the data through use of OS.

Through investigating these factors a more clear picture of the performance of the approach will be be provided. Without such a investigation the method contains could produce capable solution just as likely as poor solutions. Worse still, the setup may produce poor solutions more often than capable solutions. Once a more concrete understanding is developed of the different factors and the performance of the algorithm accordingly the research question can be answered as to whether it is possible to predict changes within a project using the commit data.

5.2.1 Prediction Features

The experimental design to allow for the predictions made on historical data to be tested with available data. Therefore within the data collected for the project the predictions must be made for values that are already known to allow for verification. Therefore the experimental sampling would build off of the prediction sampling outlined in section 4.1. A second region defined as the prediction sample range. The Figure 5.1 outlines the updated layout. The size of the training range ($|s_t|$) and the size of the prediction ($|s_p|$) are able to be different sizes, however for each experiment they remain the same ($|s_t| = |s_p|$).

The sample range is taken from the current commit c_i to c_{i-g-m} in the case that $i > m$. m denotes the size of SWR in commits and g is the number of commits the change is predicted within. For example if the model is predict a change that occurs within the next 5 commits ($g = 5$) and $m = 30$ then Figure 5.1 shows how the data would be sampled. The training sample would be where data would be collected from to train the model. The prediction gap is to account for the data sampling calculating whether methods at commit 40 will have a change within the

next 5 commits. Therefore to properly test it on data that is not used as part of the testing model the offset is needed. The SWR for the testing data set is labeled as the *Testing Sampling*.

The sliding window factor is one of core aspects related to extracting samples from the data set. When using the sliding window to sample the data the data is divided as shown in Figure 5.1. The training sample is where the training data set is sampled from. The testing sample is where the testing data is sampled from.

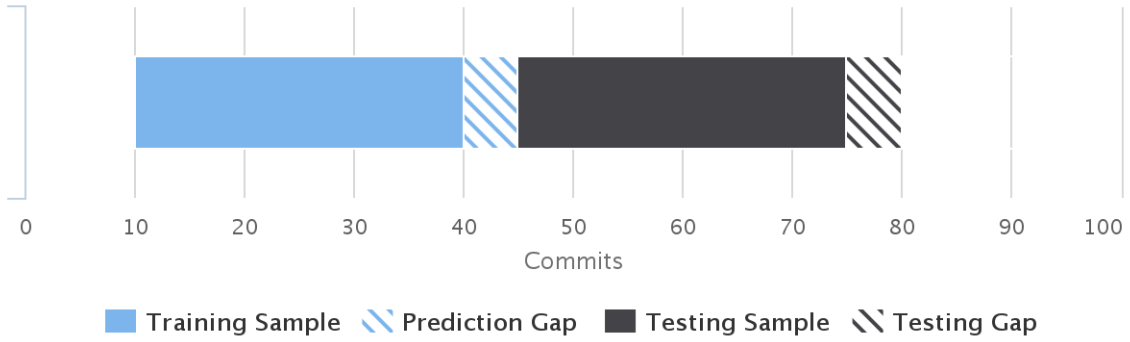


Figure 5.1: Sampling Window Layout

A data set with an extended sampling range will extend the sampling range beyond the original size for either the training sample or the testing sample. The training range can be expanded to include earlier samples to increase the sample space.

The training and testing sampling range are defined as the number of commits from which the samples can be taken. In Figure 5.1, both the training and testing sample ranges are set to 30 commits. These two values can differ from one another but tend to be kept the same for most of the experiments.

As discussed in section 4.1, sample biasing can cause the distribution to favor the selection of one category over another. Undersampling and OS can prevent the model from simply classifying all samples as one category or the other.

For each project data set there are numerous windows that be can be used. The

window number is setting which window is used broadly mapping to the position within the data set that the model will be trained on and then predicted on. In Figure 5.1 the *current commit* is located at 45. This is the point from which predictions will be made after. The gap preceding the starting point is is 5 commits long and is followed by the sample window for the training data which is 30 commits long. To calculate the window offset simply using the starting position (p_s), the gap length (g), and the SWR can be calculated in Equation 5.1. Therefore in this case the window offset would be $45 - 5 - 30$ which is 10.

$$wo = p_s - g - swr \tag{5.1}$$

Finally, the last factor of note is the parameters used to configure each prediction method. RF use a single parameter, the size of the forest. SVM meanwhile uses two parameters; C and gamma. Picking the most suitable parameters is ideal to achieve good performance from the prediction model. For SVM a grid search technique is provided by the developers of the libsvm source for optimizing the parameters. For RF, the size of the forest will have an impact but is far more manageable since it is a single parameter. A larger number of trees in the forest will generally provide better results, but will cause the algorithm to take longer to train.

5.2.2 Prediction Performance

For each experiment where the used random sampling the experiment was performed 5 times to account for variations in the random sample. Therefore if the initial results using the first sample set were not characteristic of the full dataset then running the experiment with more random samples is more likely to represent the true characteristics of the dataset. This required taking five random samples from

each quarter, training the model and running the tests on the model to then determine the average prediction score.

The goal of the prediction methods are to provide a good prediction of whether the a given vector will fit in one category or the other. A model's prediction performance can be rated using three measures of accuracy, precision and recall. Accuracy is measured as how often predictions p_i are classified correctly where a_i represents vector v_i correct classification. The prediction accuracy ($P_{accuracy}$) can then be calculated using Equation 5.6. This simply sums up the accuracy for each vector and then divides it by the total number of vectors (where $n = |v|$).

$$tp = \sum_{i=0}^{n} \begin{cases} 1 & \text{if } p_i = a_i \ \& \ a_i = 1 \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

$$tn = \sum_{i=0}^{n} \begin{cases} 1 & \text{if } p_i = a_i \ \& \ a_i = 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

$$fp = \sum_{i=0}^{n} \begin{cases} 1 & \text{if } p_i \neq a_i \ \& \ a_i = 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

$$fn = \sum_{i=0}^{n} \begin{cases} 1 & \text{if } p_i \neq a_i \ \& \ a_i = 1 \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

$$P_{accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \times 100 \tag{5.6}$$

The precision of a model is the measure of how correct the model predicts that a change will occur when it predicts that a change will occur. Given the true positives tp, represents the number of predictions that the model correctly identified as having a change and the false positives fp is the number of times the model incorrect predicted

a change to occur when it in fact did not. The equation for calculating precision is show in Equation 5.7.

$$P_{precision} = \frac{tp}{tp + fp} \tag{5.7}$$

The recall of the model is the measure of how correct the model predicts that change will occur out of all the times changes really occurred. Again using tp as the number of true positives, and false negatives fn which is the number of times the model fails to predict that a change will occur. The recall can be calculated using the Equation 5.8.

$$P_{recall} = \frac{tp}{tp + fn} \tag{5.8}$$

5.3 Experimental Results

For each experiment all of the data used to train and test the model is collected using a Ruby script to query a PostgreSQL database. The PostgreSQL database provides the raw data which is then processed into data vectors in an acceptable form for SVM or RF. The data processing method is outlined more completely in section 4.2.

5.3.1 SVM Experiments

For this set of experiments the machine learning algorithm SVM is used to provide the change predictions. As noted in section 4.2, the implementation for SVM is a Ruby binding of the original library. The parameters used for all of the experiments with SVM are $C = 10$ and $gamma = 8$.

5.3.1.1 Window Range Experiments

In this experiment the independent variable is the size of the SWR in commits. For each variation of the SWR the performance is measured. In Table 5.7, the features used by the prediction model are outlined. Features with a mark, •, are used while those without are not. In this experiment only the sf_Δ is not used while all the rest are. Each of these features is outlined in further detail in Table 4.1.

Com	Sig	Name	f_Δ	sf_Δ	t_Δ	Length	$change_{t-1}$
•	•	•	•		•	•	•

Table 5.7: SWR Experiment Features

As noted above the independent variable for this experiment is the SWR. The remaining parameters for the experiment are constant for each test. These parameters for this experiment are outlined in Table 5.8.

Extended Window	Over Sampling	Under Sampling	Sample Rate	Window Offset	SVM C	gamma
No	No	Yes	100%	5	10	8

Table 5.8: SWR Experiment Setup

Each project was tested on using these outlined parameters for an SWR varying from 60 to 130 by intervals of 10. The results for the experiments are shown with the precision, recall and accuracy. For each graph the independent variable is the number of commits in the SWR. Y-axis is the percentage for either the precision, recall or accuracy. The complete set of experimental performance results are found in subsection A.1.1. For some projects did not have enough data to complete the entirety of this experiment. For example, smile did not have enough data to complete the trials with SWR for 120 or 130. These projects were still included and show how the method works with smaller amount of data available.

The majority of the projects using SVM did not perform well with accuracy and precision typically between 0.4 and 0.6. This project as well as others have very poor results and show the difficulty of this problem. Similarly, Figure A.3 shows low precision and accuracy while very high recall. The independent variable, SWR has very little impact on the performance for this project in particular.

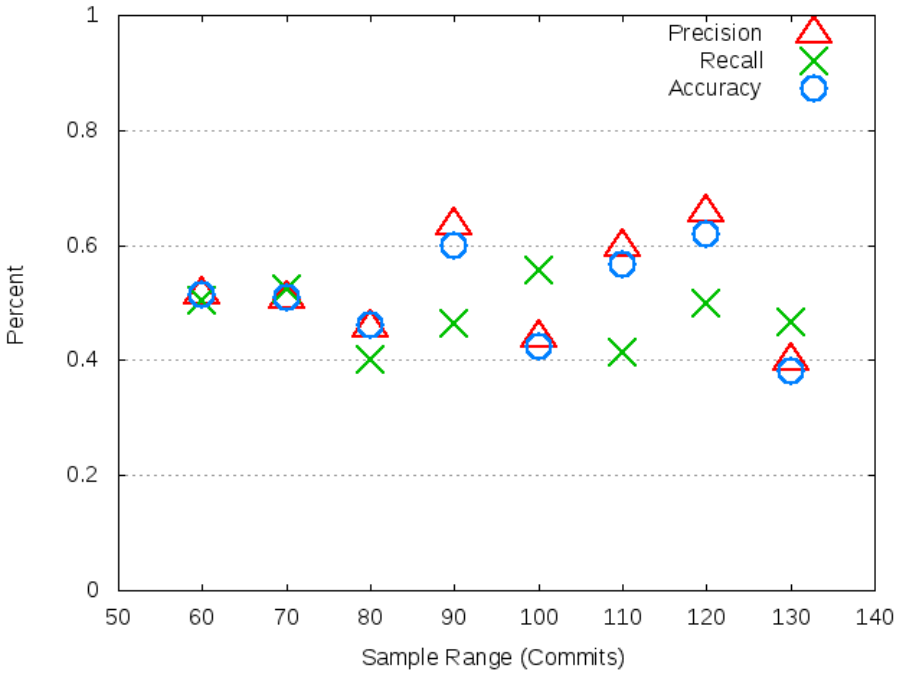


Figure 5.2: SWR for tempto using SVM

Some of the projects like http-request in Figure A.11 had a large amount of variation with the changes to the SWR. In one case http-request moderately well in SWR 80 while at 60 and 120 the accuracy and recall are 0.

In Figure A.1, the project acra is shown with the best result for SVM. When the SWR is at 70-100 the performance is high, for the other cases the performance is lower but not by a large margin. The point of interest is that recall performs well for an SWR of 100 or lower but performs worse than the accuracy and precision after 100.

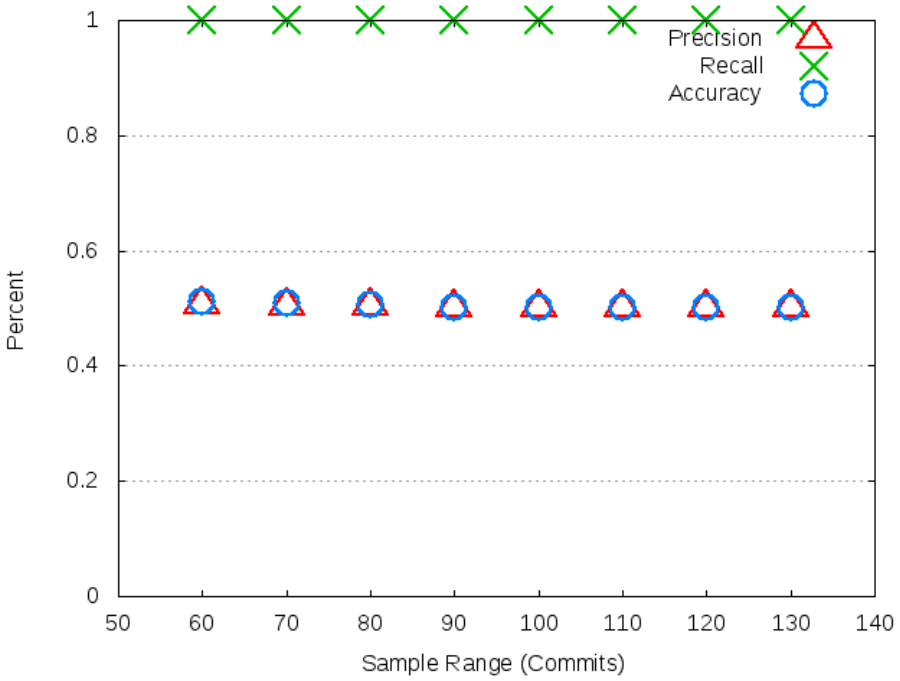


Figure 5.3: SWR for blockly-android using SVM

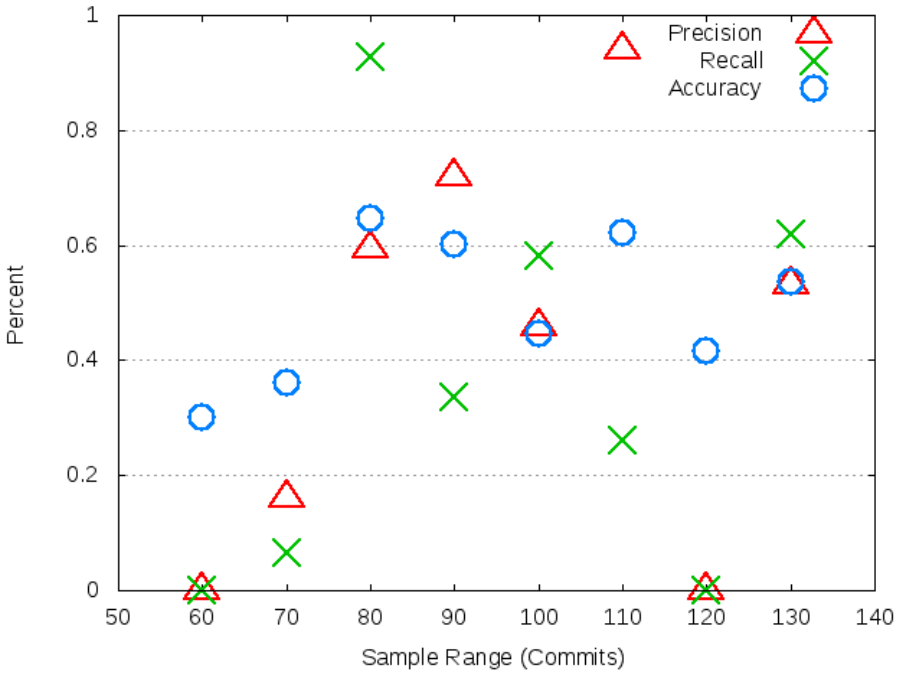


Figure 5.4: SWR for http-request using SVM

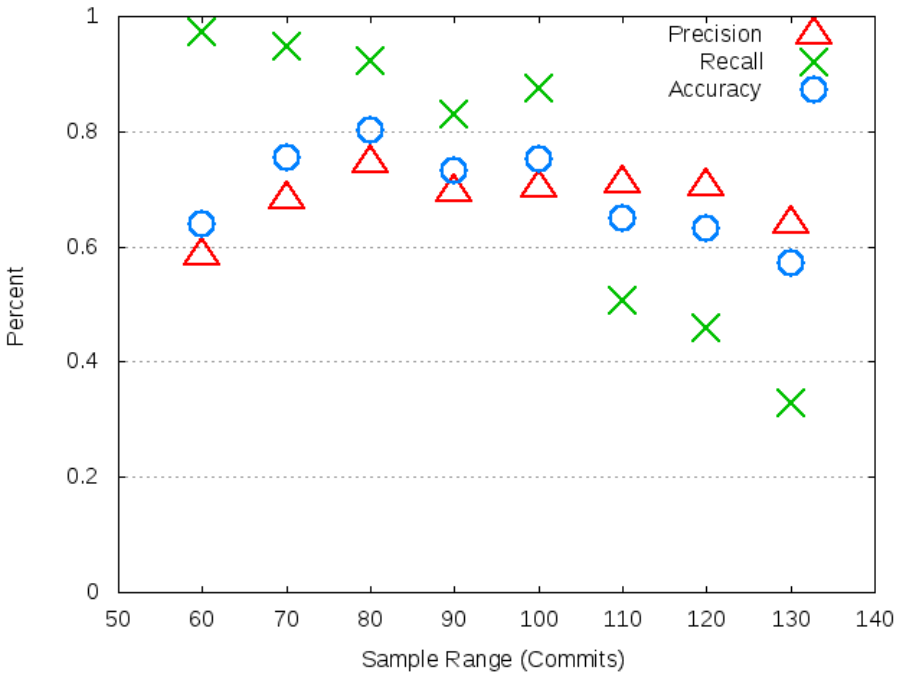


Figure 5.5: SWR for acra using SVM

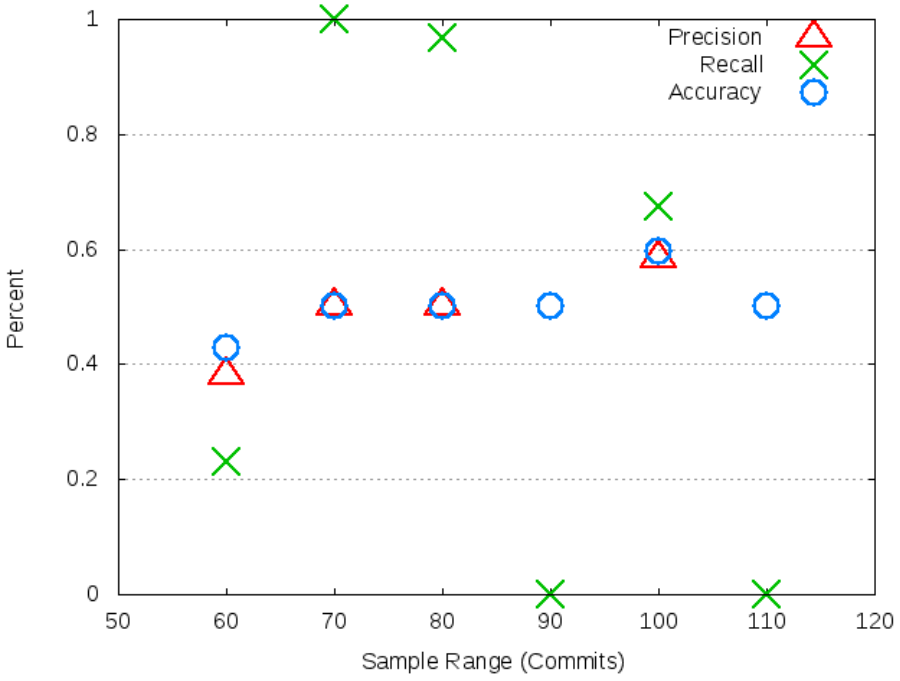


Figure 5.6: SWR for smile using SVM

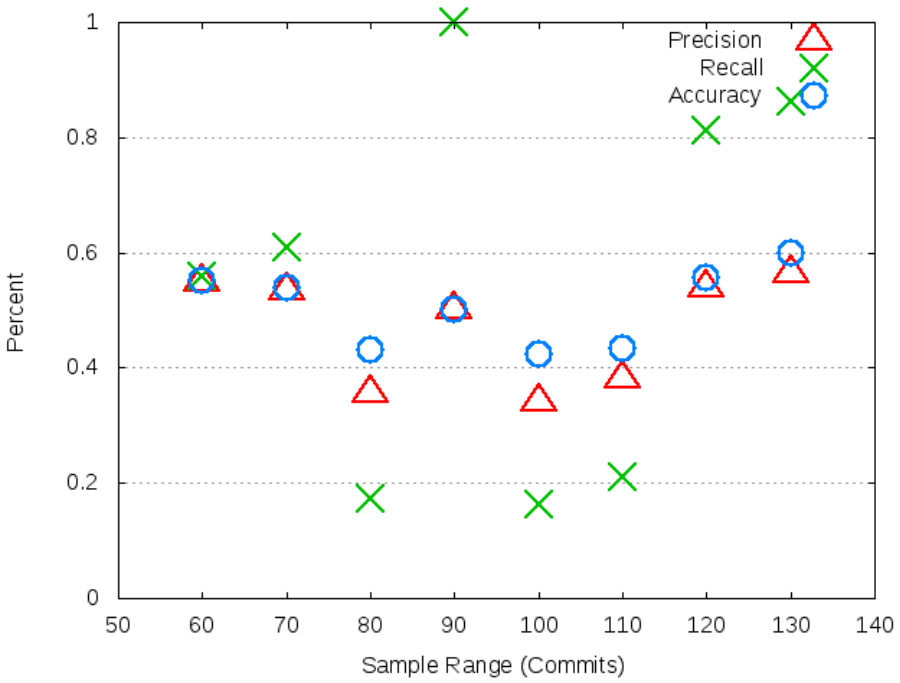


Figure 5.7: SWR for spark using SVM

Both smile in Figure A.19 and spark performed poorly with each performance measure below 0.5. In two cases (90 and 110) smile and 0 recall and undefined precision.

Overall there was no clear value for the SWR which held consistent positive results. Projects from similar groups tended to perform similarly. For example acra, dagger and ShowcaseView all tended to perform well for similar parameters.

Projects that were influenced more by SWR thus having a larger variation between values proved to have better results more often however this was not guaranteed. No value of SWR works across projects and even for projects that worked the correct value had to be found in order to obtain good results.

Extended Window	Over Sampling	Under Sampling	Sample Rate	Window Offset	SWR	SVM C	SVM gamma
No	No	Yes	100%	5	90	10	8

Table 5.9: Feature Experiment Setup

5.3.1.2 Feature Set Experiments

This experiment uses different sets of candidate feature to test to explore the available features. The remaining variables were kept constant to allow for the candidate feature sets to be viewed in isolation. These constants are provided in Table 5.9. The value of 90 was selected for the SWR based on the value being in the middle of the range experimented on for the previous experiment. The remaining variables are kept the same as the previous experiment in subsubsection 5.3.1.1.

Feature	Com	Sig	Name	f_Δ	sf_Δ	t_Δ	Length	$change_{t-1}$
1	•	•	•	•	•		•	•
2	•	•	•	•		•	•	•
3	•	•	•	•		•		•
4		•	•	•		•		•
5	•	•	•	•				•

Table 5.10: Candidate Feature Sets

The candidate feature sets are outlined in Table 5.10. These feature sets were selected from a larger set of features outlined in section 4.1. Each set is assigned an index value to allow for easier reference later on. For the remainder of this section the experiment sets will be referenced using the assigned index. Therefore if feature set 3 is referenced then that refers to the candidate feature set in the third row. Some of the projects results are shown in figures below. The rest of this experiments performance results can be found in subsection A.2.1.

Show

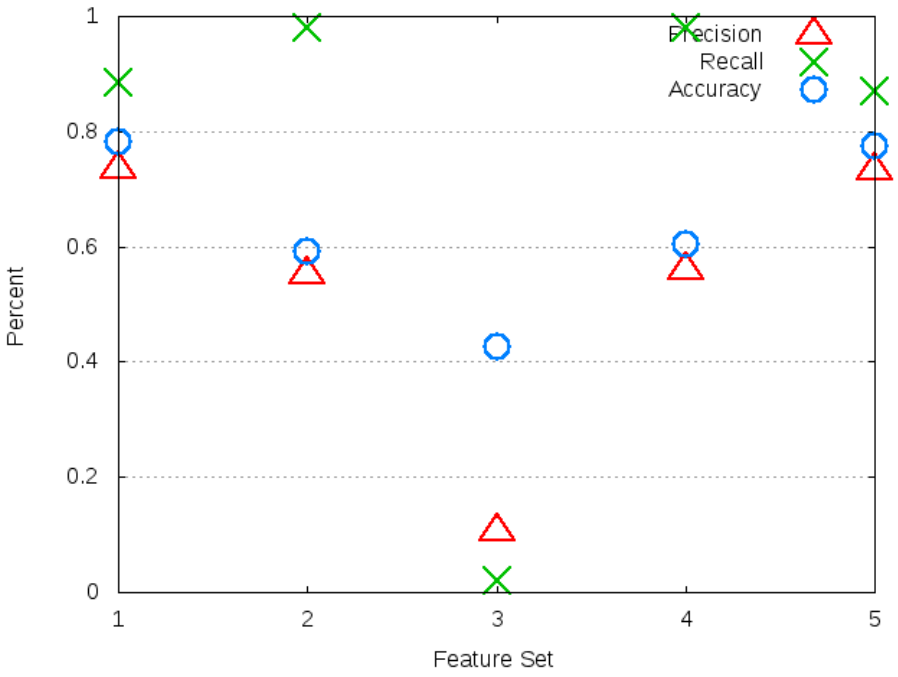


Figure 5.8: Feature for ShowcaseView using SVM

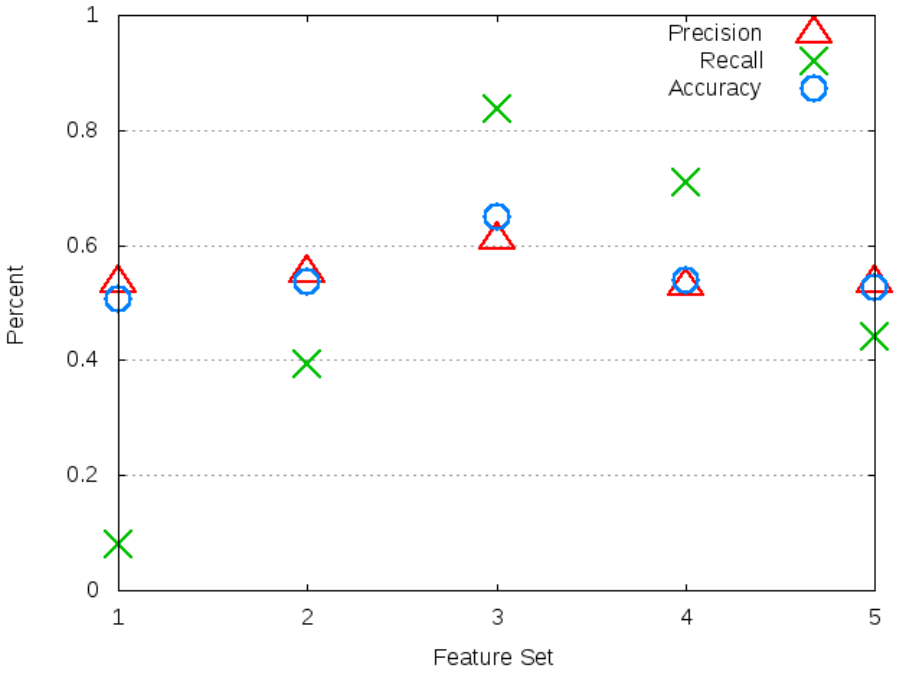


Figure 5.9: Feature for deeplearning4j using SVM

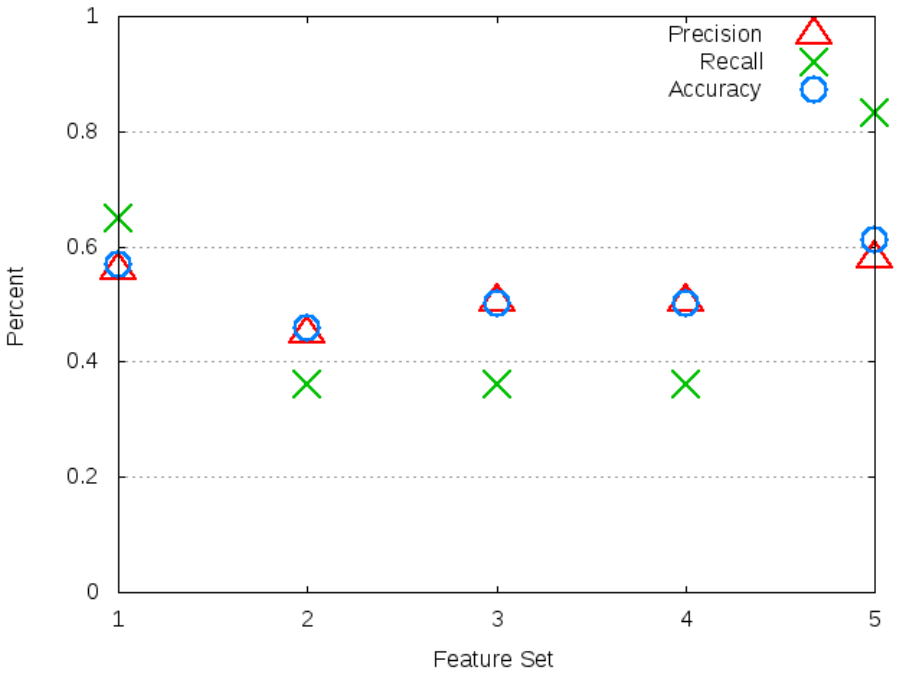


Figure 5.10: Feature for ion using SVM

The projects ShowcaseView, deeplearning4j and ion were all greatly impacted by the different feature sets. ShowcaseView in Figure A.87 performed well for feature set 1 and 5 and terribly for feature set 3. Similarly for ion in Figure A.81, feature sets 1 and 5 performed well with the rest of the feature sets performing poorly. Finally for deeplearning4j in Figure A.76, the best performance was for feature set 3 where as the remaining trails were not as good. There were a few projects like these ones were one or two of the feature sets would perform well. One that performed well for certain feature sets tended to share similar project classifications like ShowcaseView and ion do.

A lot of projects did not vary greatly for different feature sets providing similar to results to that of nettosphere in Figure A.84. All three performance measures show little variance and only small chances are present between projects. Other projects performed poorly for all the feature sets such as mapstruct in Figure A.83 which had

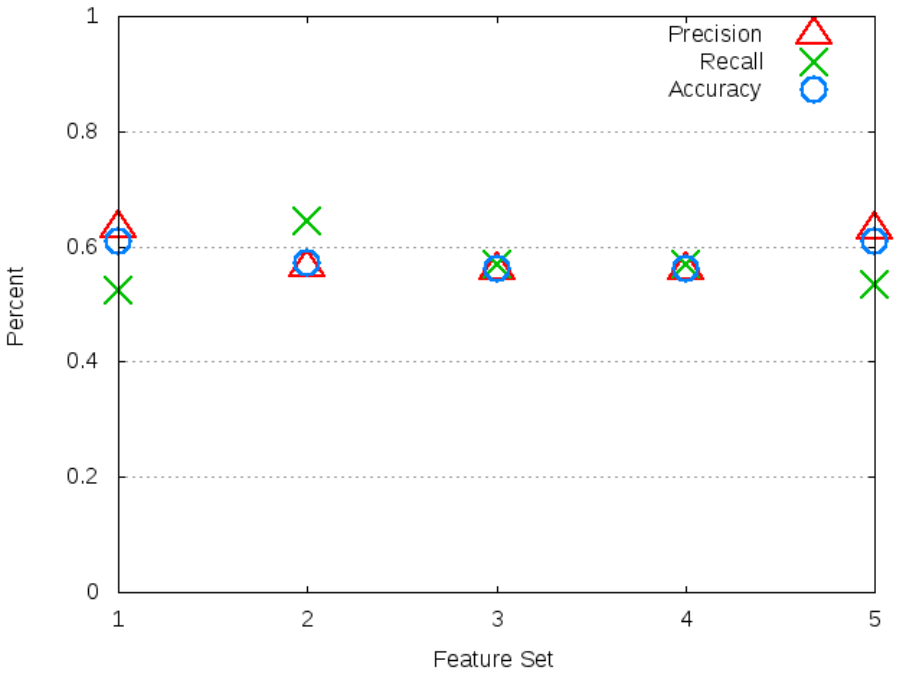


Figure 5.11: Feature for nettosphere using SVM

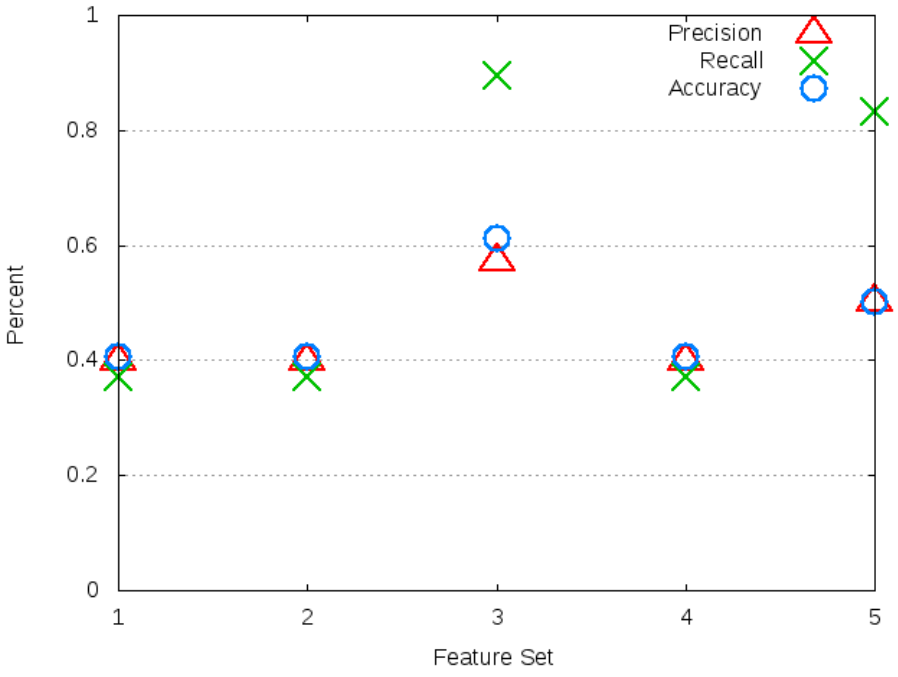


Figure 5.12: Feature for mapstruct using SVM

3 of the 5 trails score lower than 0.5 in all performance measures. The feature sets are an influencing factor on the performance of the model however no single feature set found to stand out as the ideal candidate for all projects.

5.3.1.3 SVM Oversampling Experiment

Extended Window	Under Sampling	Sample Rate	Window Offset	SVM C	gamma
No	Yes	100%	5	10	8

Table 5.11: Feature Experiment Setup

OS is a balancing technique used to increase the amount of samples available. Samples from the smaller data set are re-sampled to increase the size of the data set. While this does introduce duplicates into the model it also counter acts biasing that is present when one classification is more common then the other by a large margin. Under sampling is also used to remove excess elements from the larger set of classification. OS This is especially useful for data sets that contain a small number of samples for a particular category. In that case under sampling may limit the performance of a model by removing nearly all of the elements in the data set.

The experiment below took the best and worst trials from the previous two experiments and used OS when sampling the data. The variables that change per project are based on the previous best performance and worst performance. In Table 5.12, the best and worst SWR and feature set are provided for each project. Since each project will likely have different values of SWR and feature set the comparesion should only be made between the difference in performance for the best/worst result and their corresponding OS trail Best-O/Worst-O.

In some cases the best performing experiment may not have been entirely clear. For example with some projects having very high recall (≥ 0.9) while having lower

Project	Best		Worst	
	Feature Set	SWR	Feature Set	SWR
acra	2	80	3	90
arquillian-core	4	90	2	90
blockly-android	2	60	2	90
brave	2	130	3	90
cardslib	2	120	4	90
dagger	5	90	2	70
deeplearning4j	3	90	2	130
fresco	3	90	2	90
governator	1	90	3	90
greenDAO	4	90	1	90
http-request	2	80	3	90
ion	5	90	2	90
jadx	2	130	2	100
mapstruct	2	70	1	90
nettosphere	2	120	3	90
parceler	1	90	2	90
retrolambda	2	130	4	90
ShowcaseView	1	90	2	80
smile	2	70	2	90
spark	4	90	3	90
storm	2	100	2	110
tempto	2	120	2	130
yardstick	2	70	2	100

Table 5.12: Best And Worst Results From experiments 1 and 2 for SVM

precision and accuracy. The best trail was picked based on having the all a weighted summation algorithm outlined in Equation 5.9. Since precision and accuracy are very closely related the weight for each was 0.5 while recall was set to 1.0.

Some projects such as fresco Figure A.123 performed a little better for precision and accuracy while a little worse for precision. Unlike most projects however blockly-android in Figure A.118 experienced very little change with the introduction of OS. Finally, the majority of the experiments showed OS to provide a negative impact of the performance of the model. Both deeplearning4j in Figure A.122 and acra in

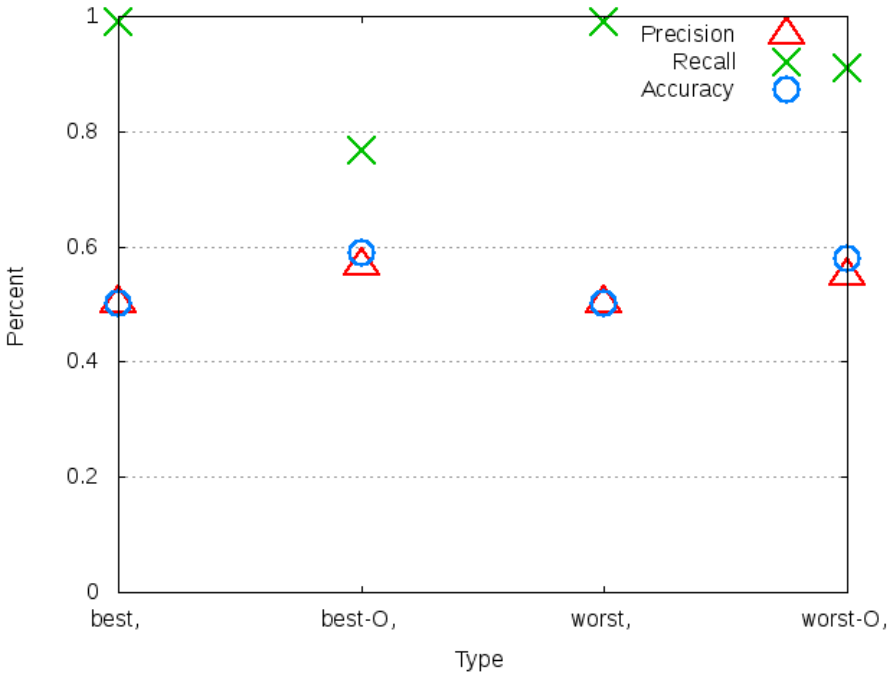


Figure 5.13: Oversampling for fresco using SVM

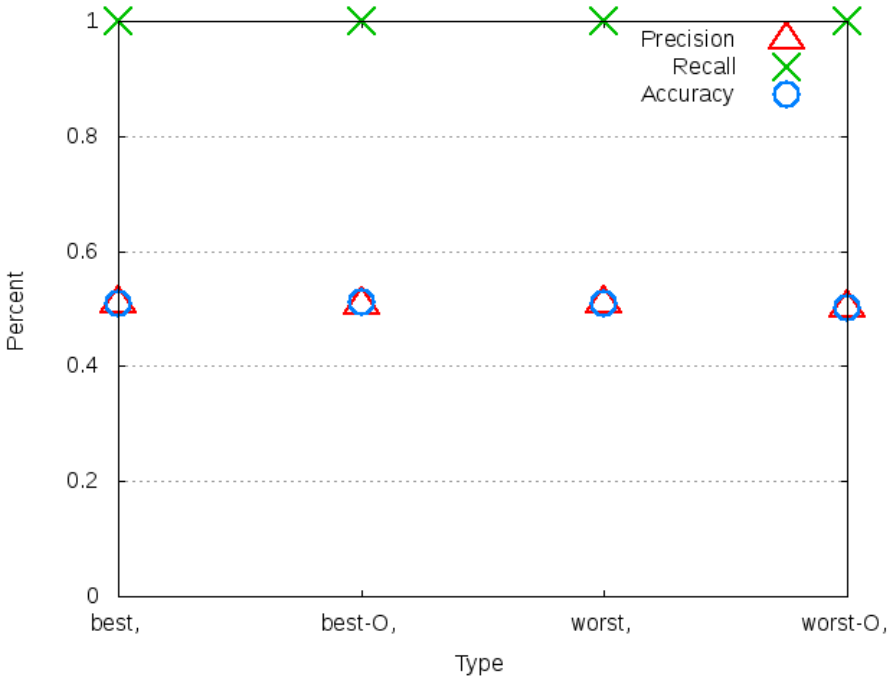


Figure 5.14: Oversampling for blockly-android using SVM

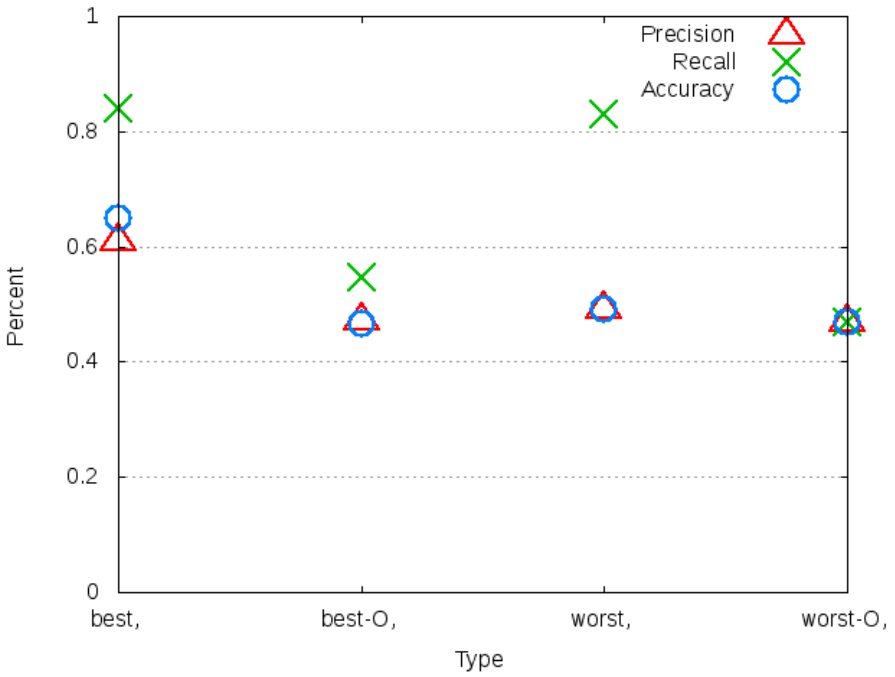


Figure 5.15: Oversampling for deeplearning4j using SVM

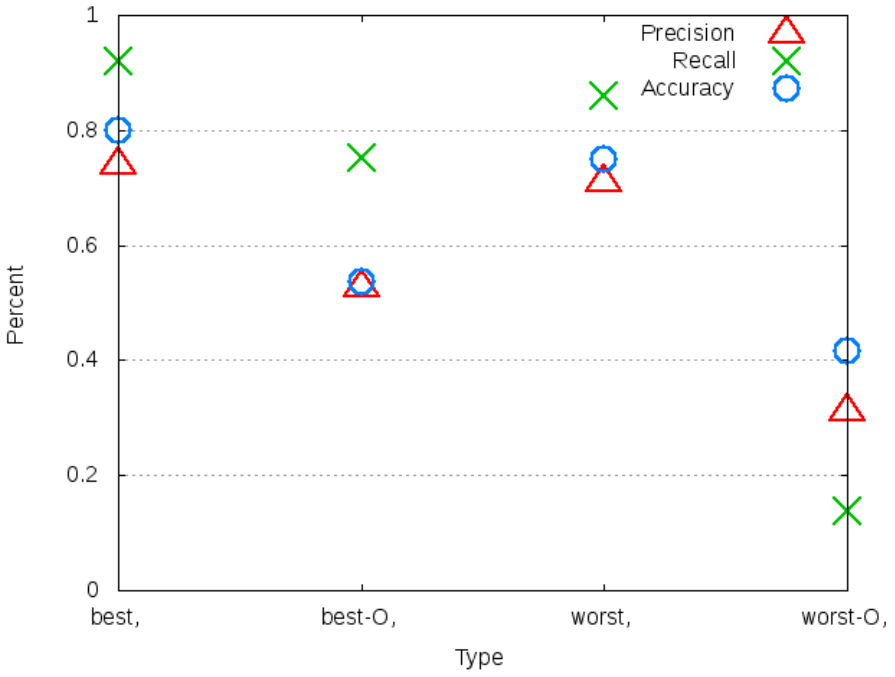


Figure 5.16: Oversampling for acra using SVM

Figure A.116 performed worse for all measures for both trails.

The overall impact of using OS for training the model proved detrimental with the vast majority of projects performing worse with the use of OS. While some projects experience increases in individual performance measures, other measures fall. Also, any performance seen by a project is minimal at best while the loss of performance tends to be substantial.

5.3.1.4 SVM Discussion

The three different experiments attempted to determine the impact of the different factors on the prediction method. The three factors that were tested are:

1. SWR
2. Model features
3. Sampling balancing

The results of the projects did not follow any common trends between projects. Similar project groups outlined in Table 5.2 do not perform similarly for the most part with the exception of the group for acra, dagger and ShowcaseView which all had an okay but for different values of SWR. The only similarity that can be found is for groups of project that performed poorly which still was inconsistent. For most projects at the very least the recall was variable. The only really exception would be blockly-android which experienced little to no variation in for each trial. Some projects saw more variation in the precision and accuracy and often had at least trail which performed moderately well.

For the second experiment, the performance results were a lot lower than the first generally. This appears to be directly related to the impact that each of these variables has on the performance of the model. Therefore the focus is placed on the

best feature set for all projects or for groups of projects. For some of the groups of projects, a certain feature set or pair of feature sets performed well for all projects in the group. For example, acra, dagger and ShowcaseView all performed well when using feature set 1 or 5. Not all of the projects performed their best using that those feature sets. However, they did perform close to their best performance. While this trend occurred for some projects it was not consistent for all projects. Similarly there was no best performing feature set for all project.

Finally of the third experiment was found to have a variable impact on the projects when sample balancing through OS was applied. In terms of performance impact, OS provide a negative impact on most of the projects experimented on. Some trials saw now change and a very small number of trails saw slight improvements to one performance measure while decreasing another performance measure. The slight improvements found from using OS were insignificant compared to the drop in performance typically experienced.

Overall SWR had the greatest impact on the performance of the prediction method. The model feature set had less of an impact and balancing the sample through OS provided a primarily negative impact. The SWR while having a larger impact on most projects, some projects were less affected. Generally variations of SWR could produce a positive results, a negative results were also present. Furthermore, no clear pattern was discovered to allow for simple configuration of the parameters to provide positive results. Therefore use of the approach with a SVM model can be beneficial but also incurs a risk associated with poor predictions.

5.3.2 Random Forest Experiments

The machine learning algorithm RF is used for the second set of experiments. RF was selected as an alternative to SVM for it's success in various data mining related

tools. The implementation of RF is in a python library *scikit-learn* which is outlined in section 4.2. Only one parameter is used for RF, the forest size, which is set to 10000 all of these experiments.

5.3.2.1 Window Range Experiments

Com	Sig	Name	f_Δ	sf_Δ	t_Δ	Length	$change_{t-1}$
•	•	•	•		•	•	•

Table 5.13: SWR Experiment Features

Extended Window	Over Sampling	Under Sampling	Sample Rate	Window Offset	RF Size
No	No	Yes	100%	5	10000

Table 5.14: SWR Experiment Setup

The independent variable for this set of experiments is the sample window size measured in commits. The feature set are outlined in Table 5.13. The features used for this experiment is the same as the first SVM experiment feature set.

The parameters for this experiment are outlined in Table 5.14. The only difference between the parameters used in this experiment and the parameters used in the SVM experiment one is the RF specific parameters. This allows for a fairly clear comparison between these two methods with the given independent variable, the SWR. The experiment was conducted on all 23 projects collected and examples were are discussed in more detail in this section. The remainder of the projects performance results are outlined in subsection A.1.2.

Each projects experimental results are accompanied with a second figure that outlines the importance for each feature set variable in the creation of the prediction model. The importance of a feature only identifies how influential that feature was

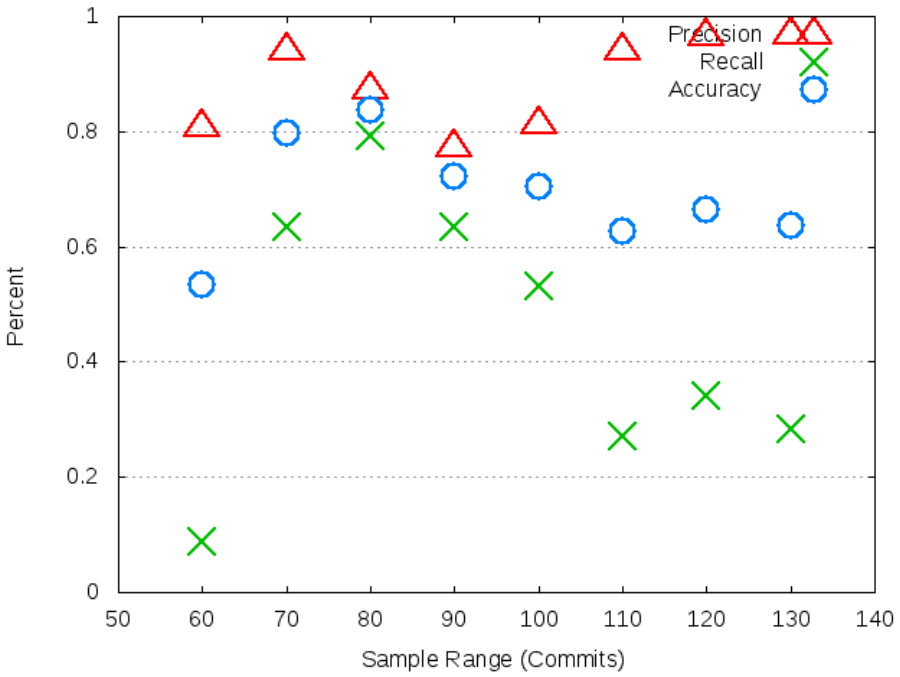


Figure 5.17: SWR for http-request using RF

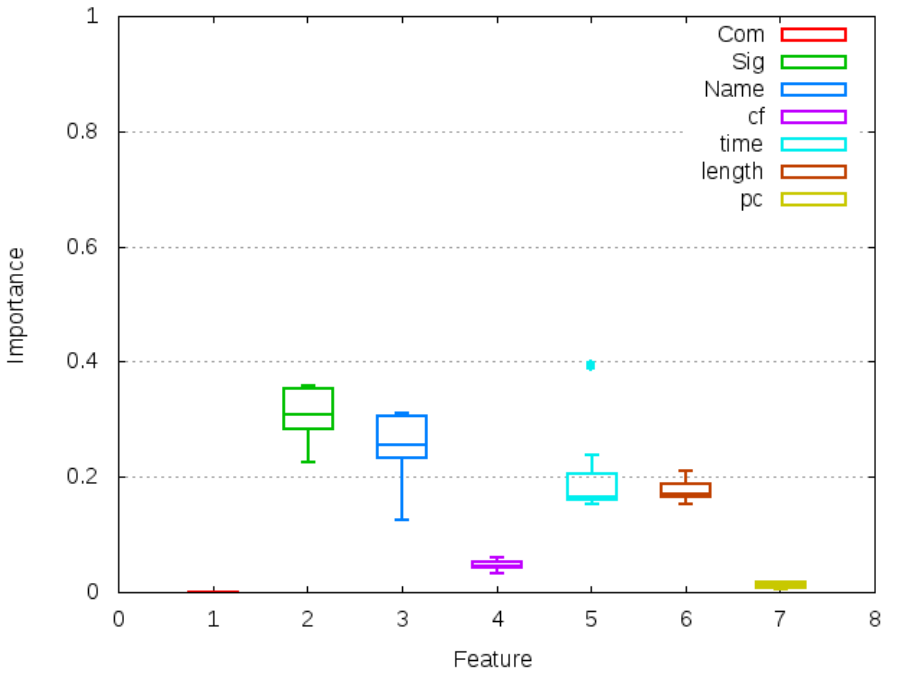


Figure 5.18: Feature Importance SWR for http-request using RF

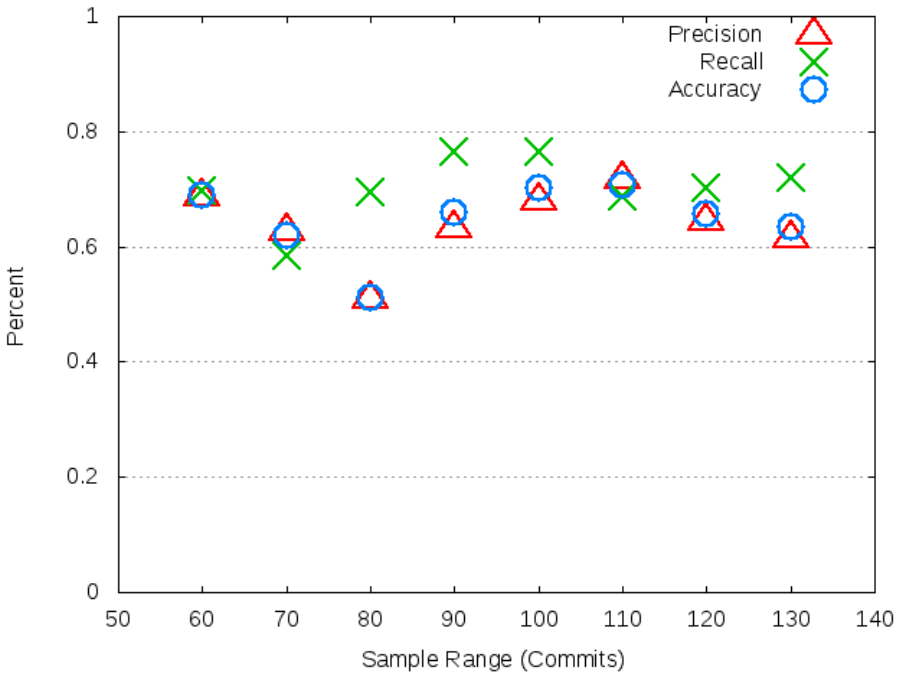


Figure 5.19: SWR for dagger using RF

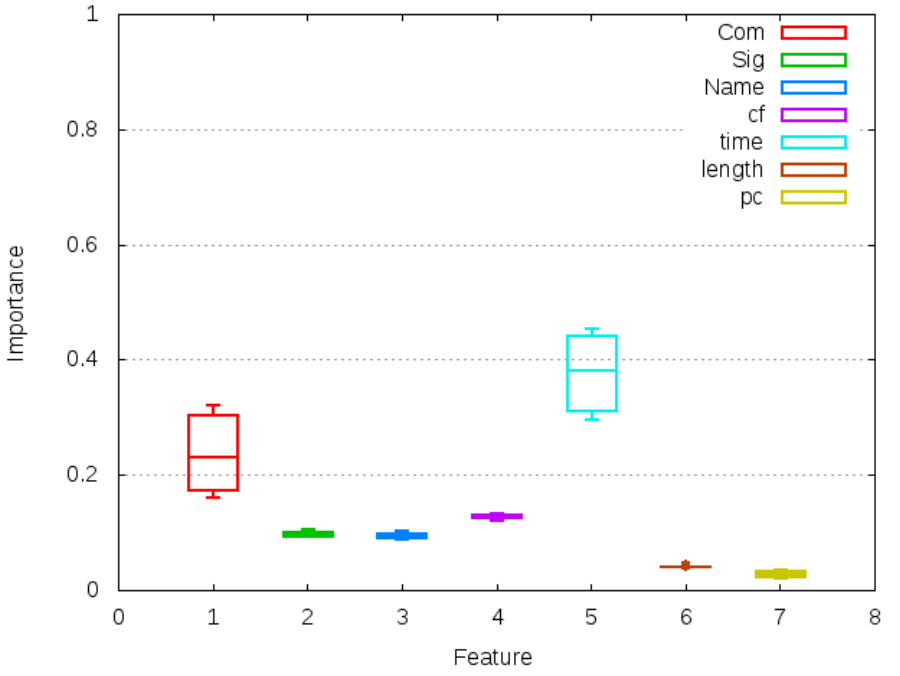


Figure 5.20: Feature Importance SWR for dagger using RF

to the prediction and therefore necessitates the context of the results. So if a project performs poorly in predicting the most influential features are more likely to not as useful for predictions with the project. Likewise, if a project performs well the corresponding feature importance can indicate highly influential features that helped produce positive results.

Even in the case of where a project performs well the feature set used may not be generalizable to other projects. For example http-request in Figure A.45 performs well with a SWR of 70 - 90 and places high importance on *Sig*, *Name*, *time* and *length*. Alternatively, dagger in Figure A.35 performs moderately well with an SWR of 60 and 100-110 while placing a higher importance on *Com* and *time*. Even more interesting is that http-request placed nearly 0 importance on *Com* while the same feature ranked second for dagger.

The performance of each project varied, with a few projects performing well for some SWR like http-request in Figure A.44, dagger in Figure A.34 and ShowcaseView in Figure A.58. The impact of the changes to SWR is clearly visible as some trails perform poorly, while others perform a lot better. For example in ShowcaseView, use of a SWR of 100 or higher provides good performance but below the performance is a lot lower. For some projects such as jadx in Figure A.48 the SWR had less of an impact causing little variation between precision and accuracy while offering only slight changes with the recall.

Some projects experienced very little impact from variations of the SWR for the performance. For example, storm in Figure A.64 had higher recall but lower precision and accuracy. Other factors may provide more influence such smaller or larger values of SWR however those were outside the scope of the experiment. Finally, other projects did not perform as well but experienced some variation to precision, recall and accuracy. One such project would be Figure A.54 which had all three measures

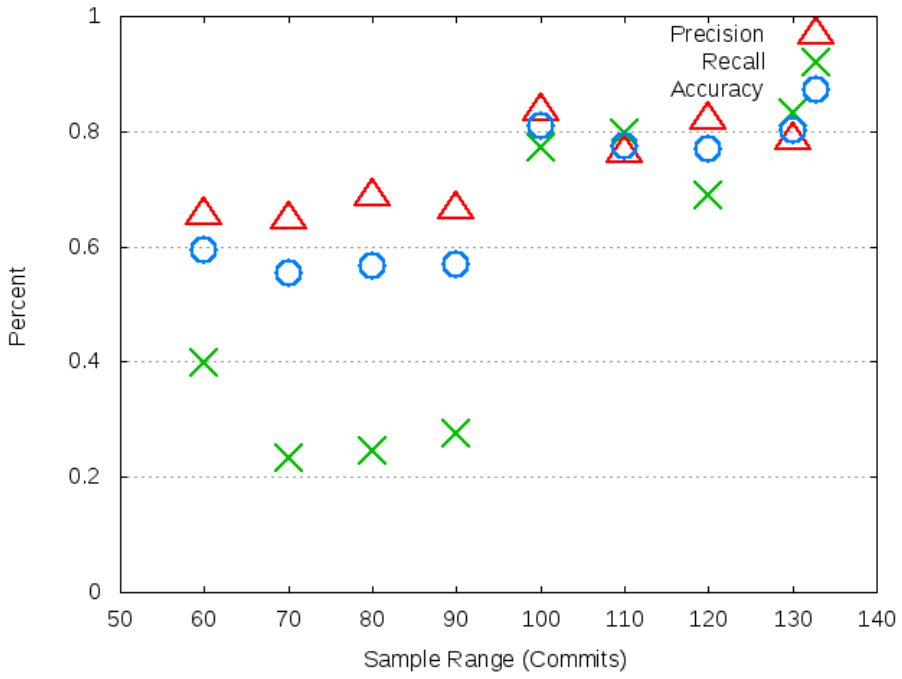


Figure 5.21: SWR for ShowcaseView using RF

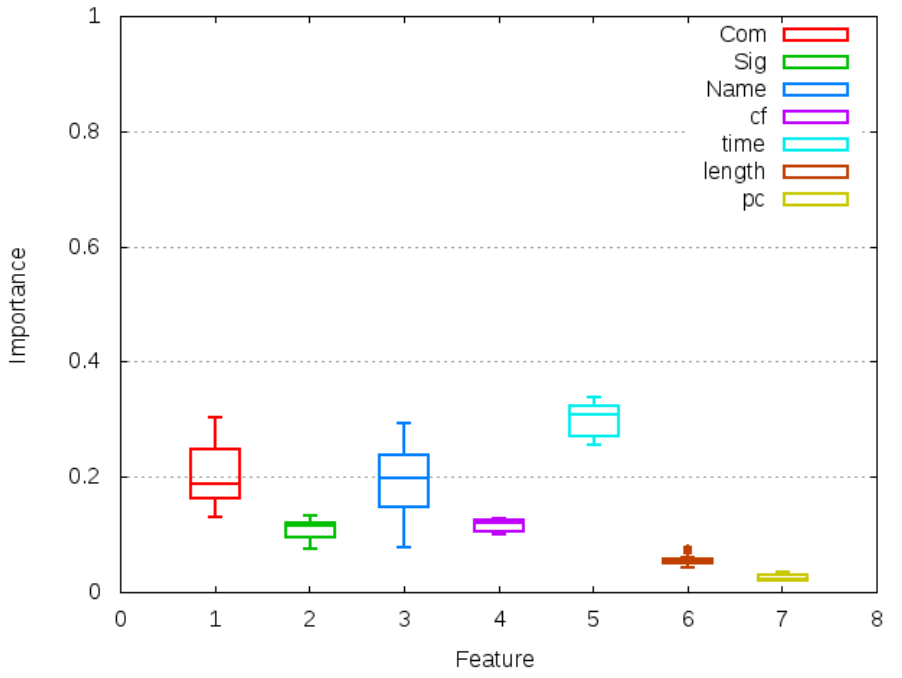


Figure 5.22: Feature Importance SWR for ShowcaseView using RF

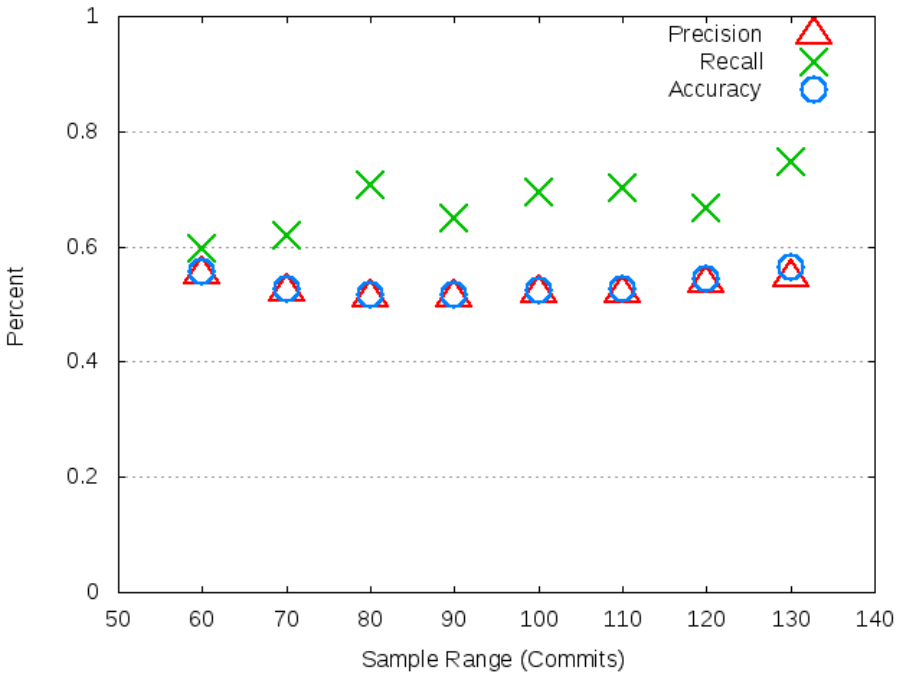


Figure 5.23: SWR for jadx using RF

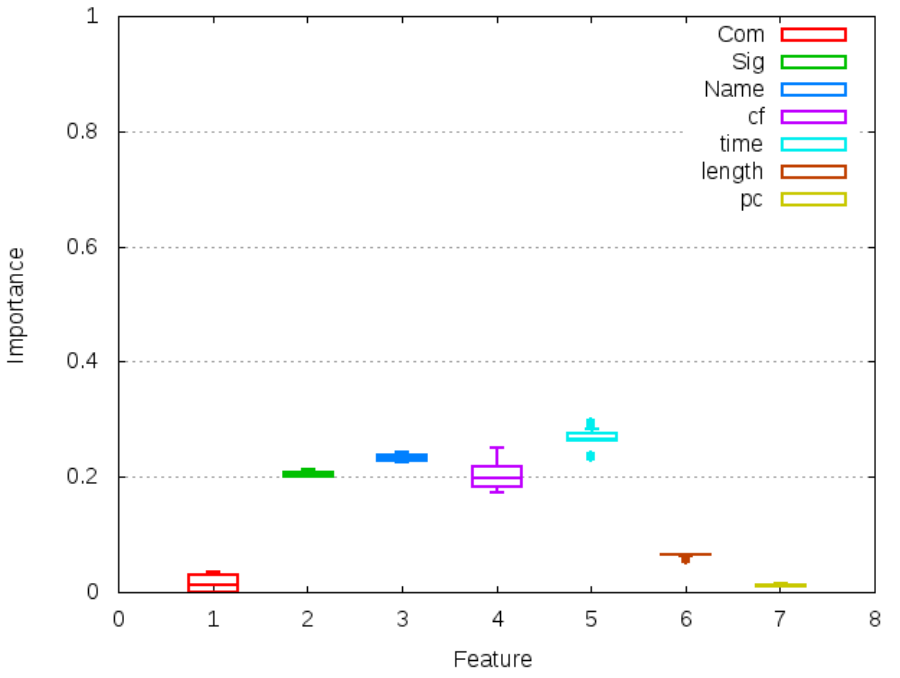


Figure 5.24: Feature Importance SWR for jadx using RF

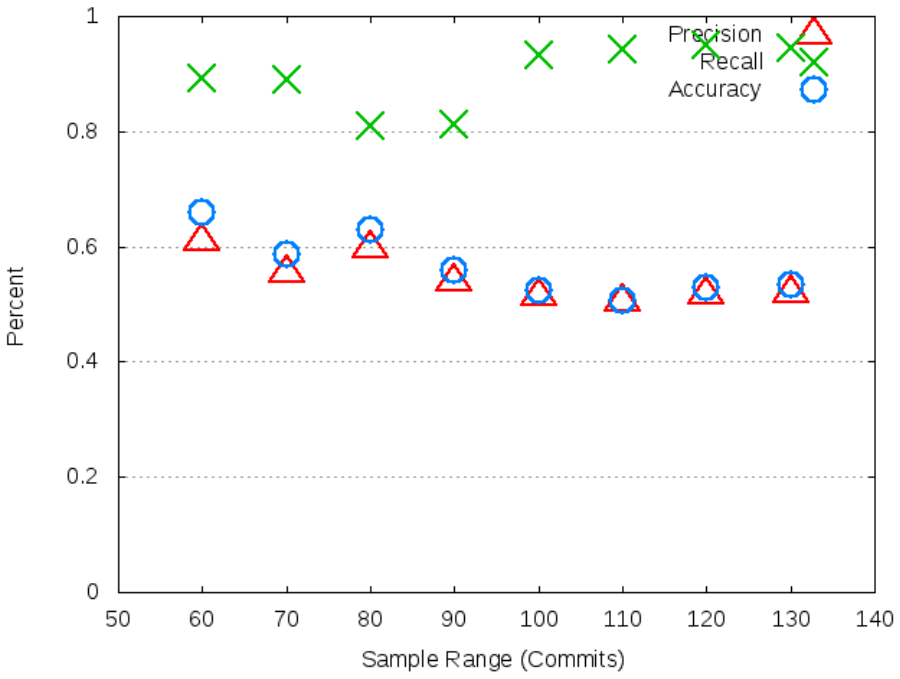


Figure 5.25: SWR for storm using RF

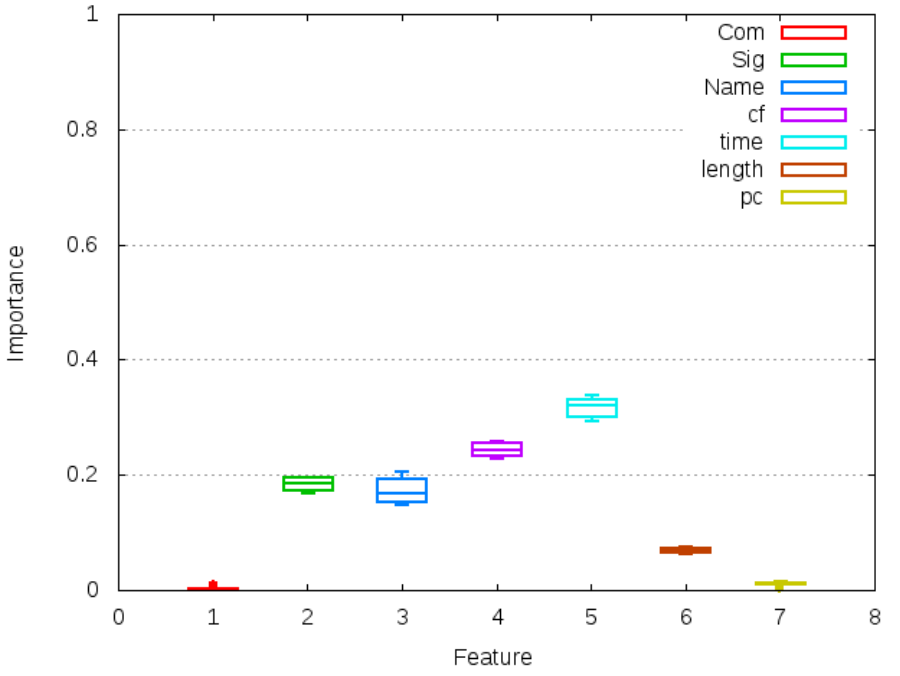


Figure 5.26: Feature Importance SWR for storm using RF

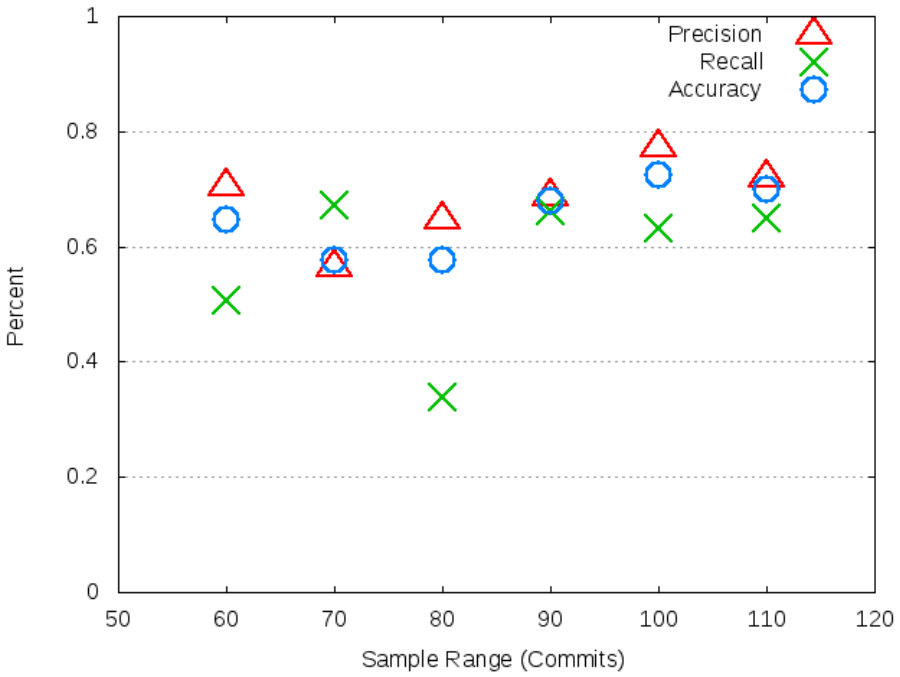


Figure 5.27: SWR for parceler using RF

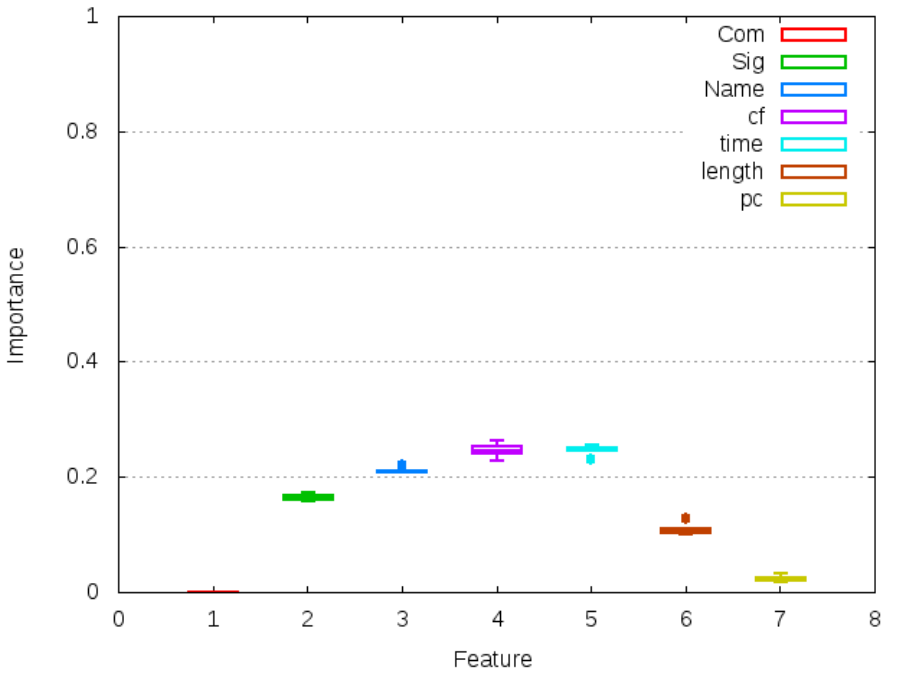


Figure 5.28: Feature Importance SWR for parceler using RF

fairly close together for most trails but because of the size of the project could not supply sufficient data for a SWR of 120 or 130. The importance of both storm and parceler was similar to that of http-request however as noted neither managed to perform as well as the best performance from http-request.

5.3.2.2 Feature Set Experiments

Extended Window	Over Sampling	Under Sampling	Sample Rate	Window Offset	SWR	RF Size
No	No	Yes	100%	5	90	10000

Table 5.15: Candidate Feature Experiment Setup

Feature	Com	Sig	Name	f_Δ	sf_Δ	t_Δ	Length	$change_{t-1}$
1	•	•	•	•	•		•	•
2	•	•	•	•		•	•	•
3	•	•	•	•		•		•
4		•	•	•		•		•
5	•	•	•	•				•

Table 5.16: Candidate Feature Sets

Similar to the experiment using a SVM in subsubsection 5.3.1.2. The experiment parameters are outlined in Table 5.15. The candidate features are likewise outlined in Table 5.16. Each set is assigned an index value to allow for easier reference later on in this section. The candidate feature set will be referenced by the index assigned in the plots and discussions related. The candidate feature sets were used experimented on with each project which are discussed below. The following results are highlights of the larger set of experiments conducted on each project. The remainder of the results for this experiment are outlined in subsection A.2.2.

The project to perform the best in this experiment was ShowcaseView in Figure A.110 which performed best for feature sets 1 and 5 and had minimal difference

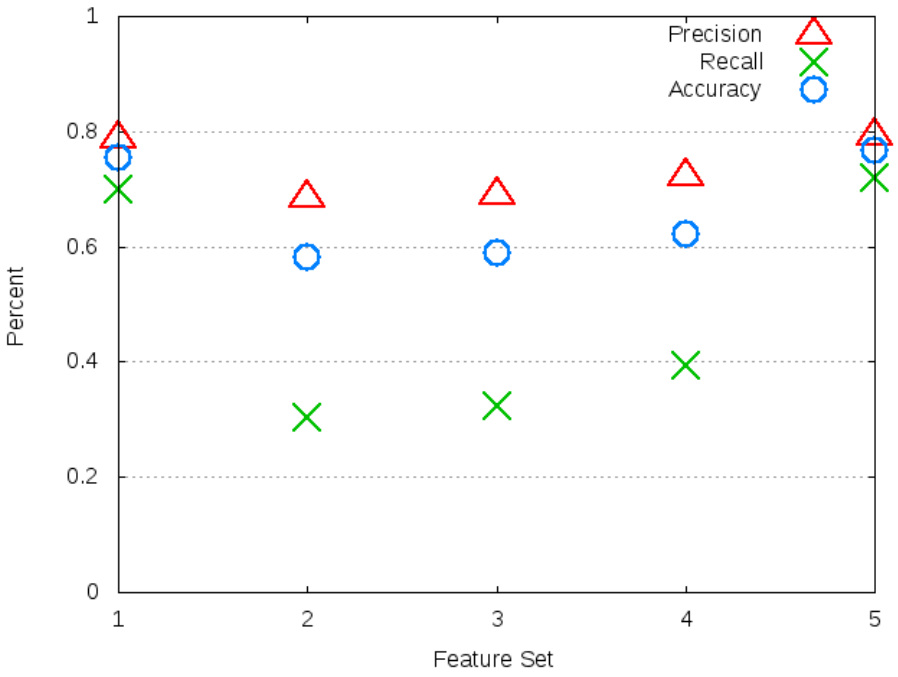


Figure 5.29: Feature for ShowcaseView using RF

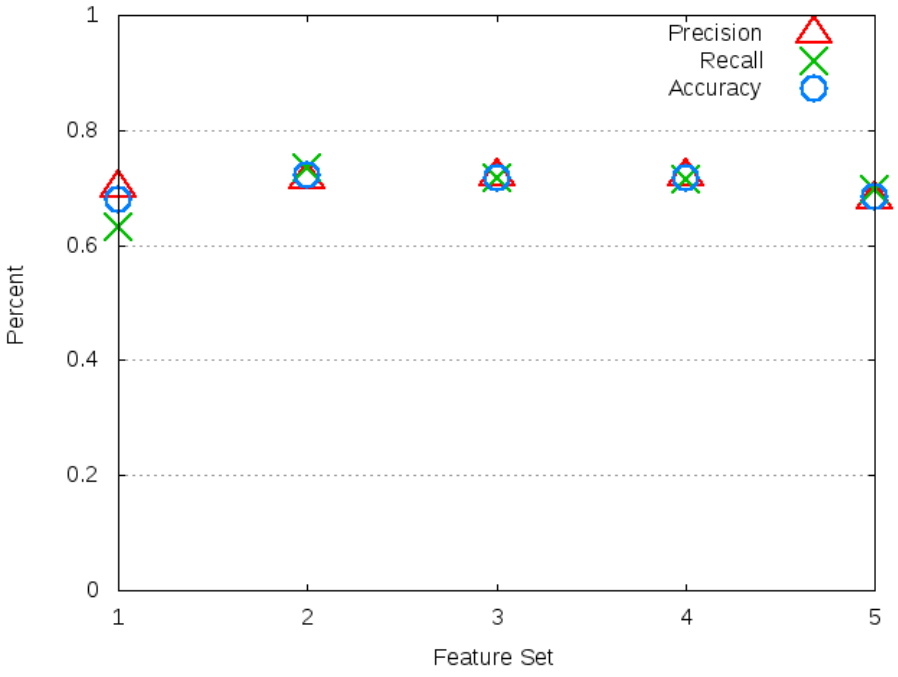


Figure 5.30: Feature for ion using RF

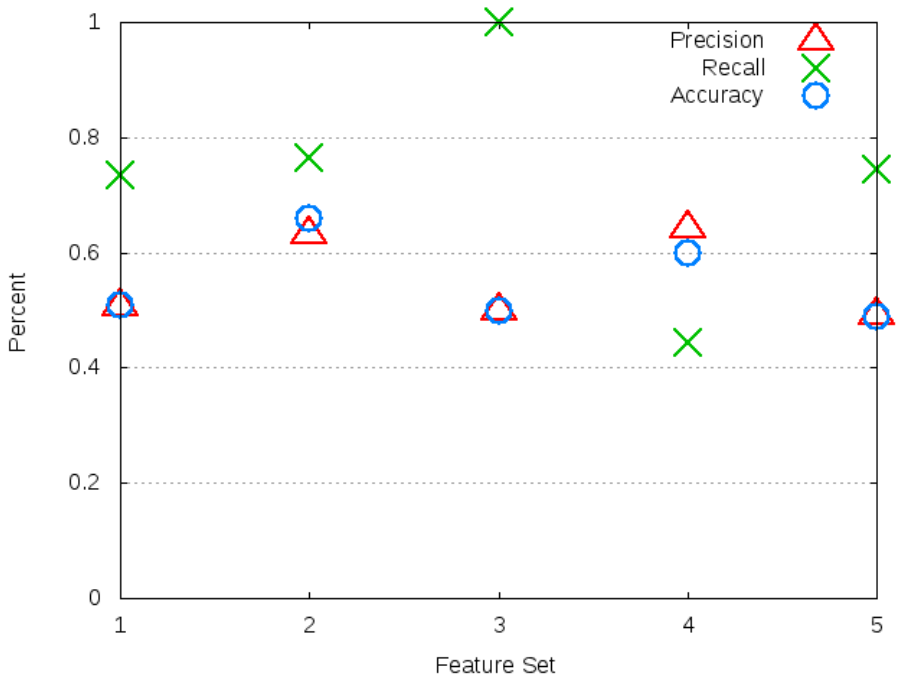


Figure 5.31: Feature for dagger using RF

between the performance of feature sets 2, 3 and 4. Typically most projects performed well with two feature sets or more which most likely is related to the similarity in the feature sets tested. A few of the successful projects, such as ion in Figure A.104, performed well consistently for each feature set. Finally, other more successful projects saw dramatic variations between the different feature sets. For example, dagger in Figure A.98, performs well in the second feature set and poorly in the rest. The recall is especially volatile at dropping below 0.5 for feature set 4.

The majority of the projects experimented on saw little difference for each feature set. For cardslib in Figure A.97, the precision and accuracy stay right above 0.5 while the recall dips below 0.5 for feature set 4. The performance is not great and overall the impact of the different feature sets appears quite low for this project. Alternatively, governator in Figure A.101, is one of the few projects to perform well for precision while low for accuracy and very low for recall. Each feature set again has only a small

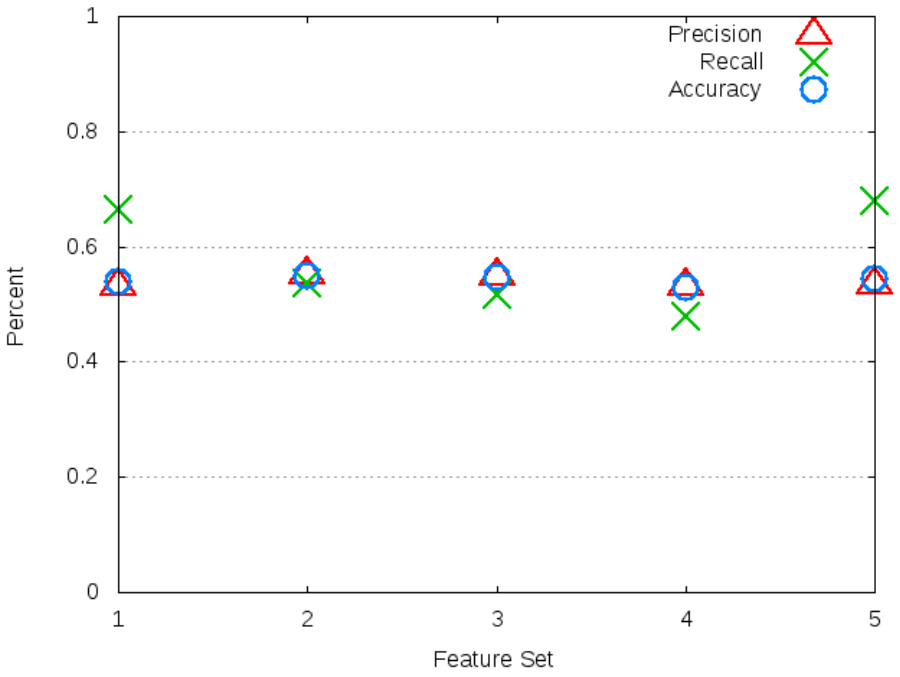


Figure 5.32: Feature for cardslib using RF

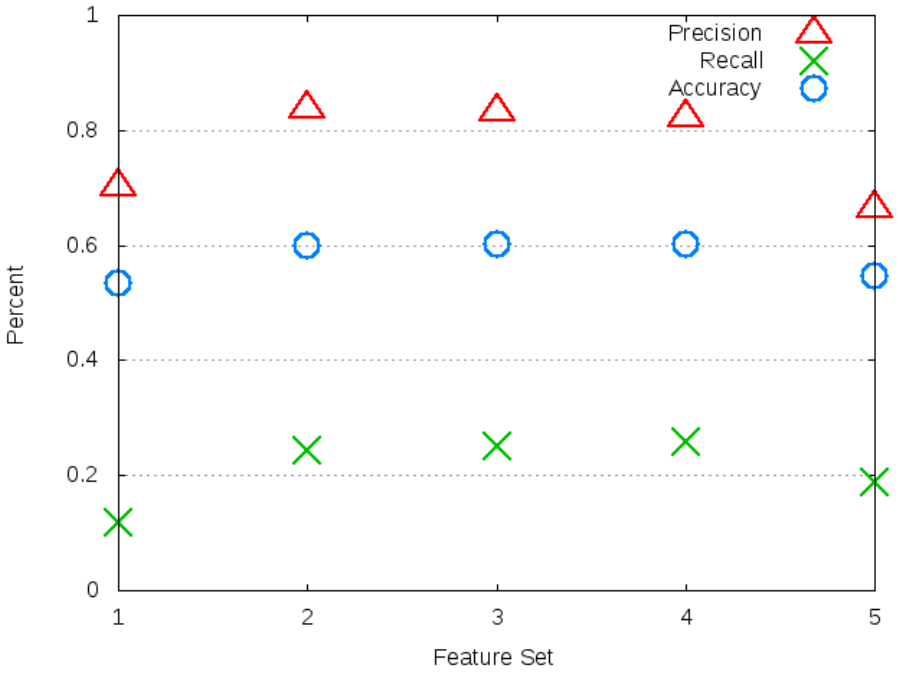


Figure 5.33: Feature for governator using RF

impact on the performance with the project performing very poorly for every trial. Overall the results for this experiment were mixed since the impact of the feature set at least of these three features is less significant that the SWR.

5.3.2.3 Oversampling Experiment

Extended Window	Under Sampling	Sample Rate	Window Offset	RF Size
No	Yes	100%	5	10000

Table 5.17: OS Experiment Setup

This experiment builds on top of the previous two experiments and shares a very similar setup to those experiments.The experiment parameters are outlined in Table 5.17. The best and worst trails for each project are taken from the previous two experiments. The value for SWR and the feature set used were recorded in Table 5.18 for each project for the best and worst performance of the RF model. The experiment applies OS the best and worst trials to compare the performance of the model with and without the use of OS.

The result of a trail with OS are represented in the figures by either *best-O* or *worst-O*. The results without OS from the previous experiment are represented with *best* and *worst*.

The results for the use of OS on the best and worst trails from each project provide to have little effect on the results. In rare cases such as dagger in Figure A.144, the performance marginally improved for the some measures while decreasing for others. However dagger also performed worse when OS was used on the worst trial. Similarly yardstick in Figure A.161, performed slightly better for best-O and dramatically worse for worst-O.

Project	Best		Worst	
	Feature Set	SWR	Feature Set	SWR
acra	2	60	5	90
arquillian-core	3	90	4	90
blockly-android	2	90	1	90
brave	2	110	4	90
cardslib	2	100	4	90
dagger	3	90	2	80
deeplearning4j	2	70	2	80
fresco	2	60	2	90
governator	2	60	2	90
greenDAO	2	60	2	100
http-request	2	80	4	90
ion	2	90	1	90
jadx	2	130	2	70
mapstruct	1	90	5	90
nettosphere	2	110	2	90
parceler	3	90	1	90
retrolambda	2	120	2	90
ShowcaseView	2	130	2	90
smile	5	90	2	100
spark	2	110	2	80
storm	2	60	2	90
tempto	2	120	2	130
yardstick	2	70	2	60

Table 5.18: Best And Worst Results From Experiments 1 and 2 for RF

The majority of the projects however saw no improvement or a slight decrease in performance. For example arquillian-core in Figure A.140 performed around the same for best-O and slightly worse for worst-O. Generally the differences if any were very small for most. Finally, some projects performed a lot worse for their worst-O trial including greenDAO in Figure A.148. Overall the use of OS on either the best or worst case cause little difference. The only major difference was for worst-O which typically performed poor in compared to the trial without the use of OS.

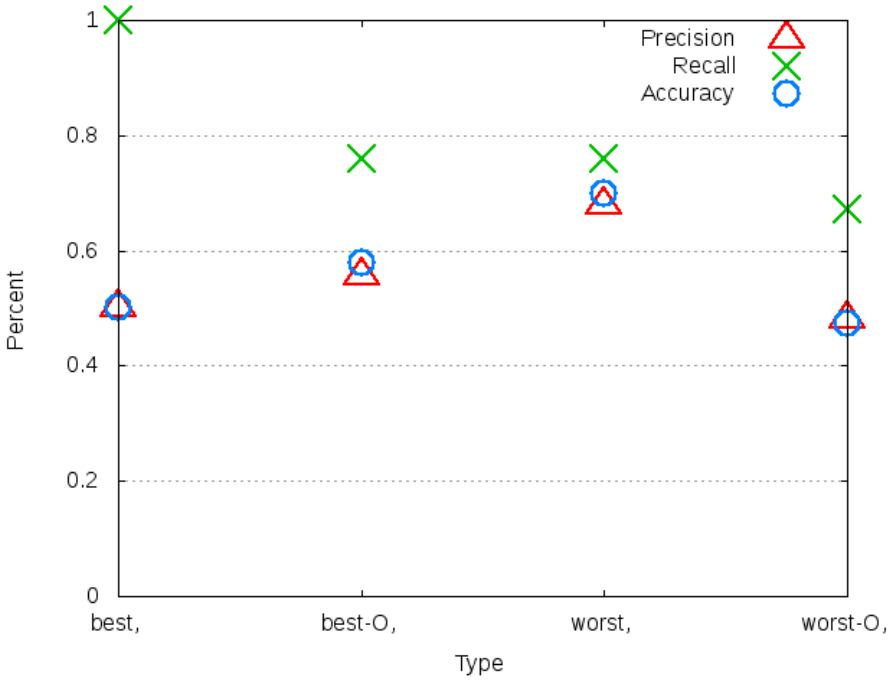


Figure 5.34: Oversampling for dagger using RF

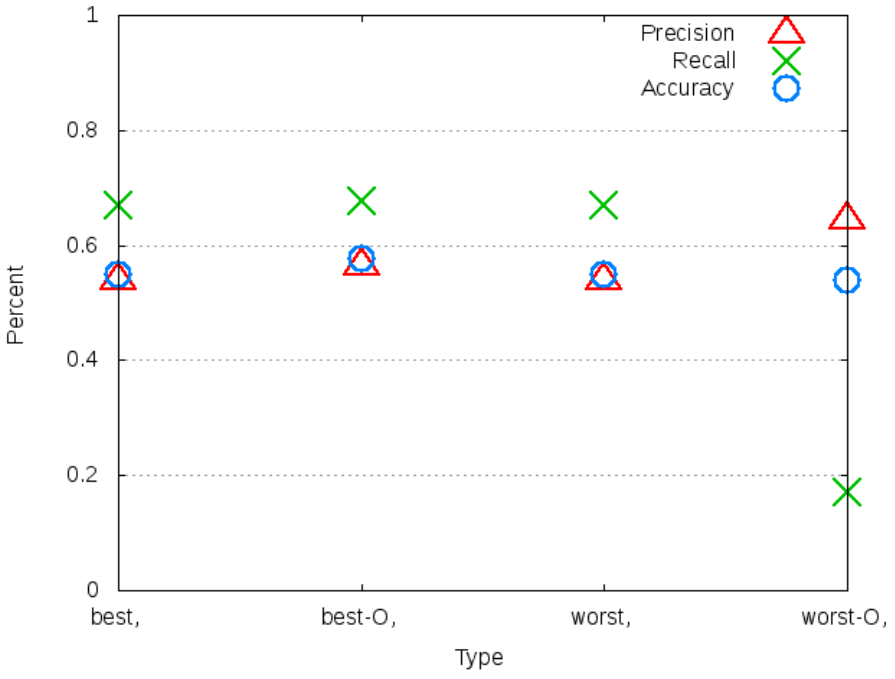


Figure 5.35: Oversampling for yardstick using RF

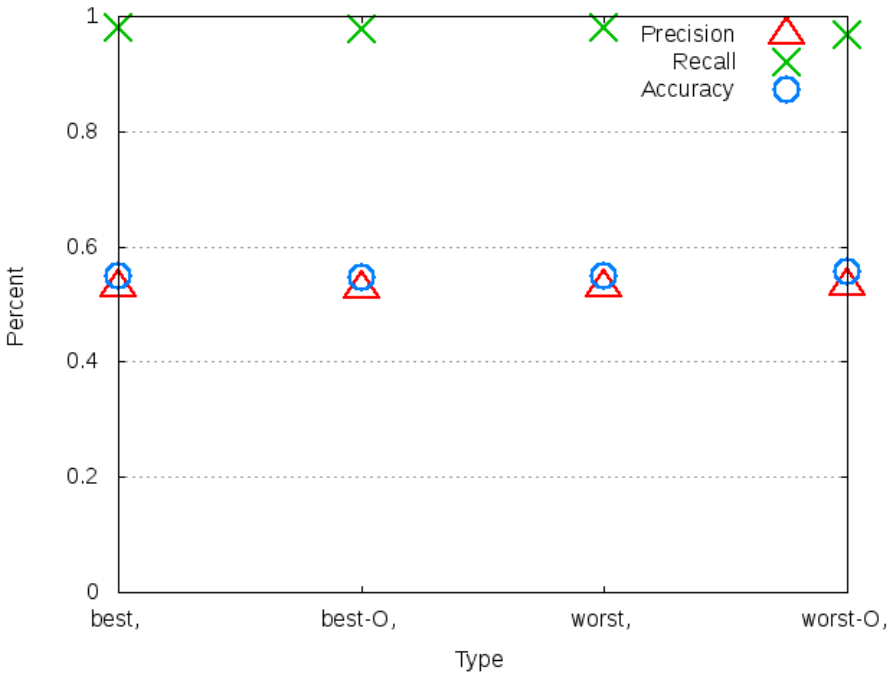


Figure 5.36: Oversampling for arquillian-core using RF

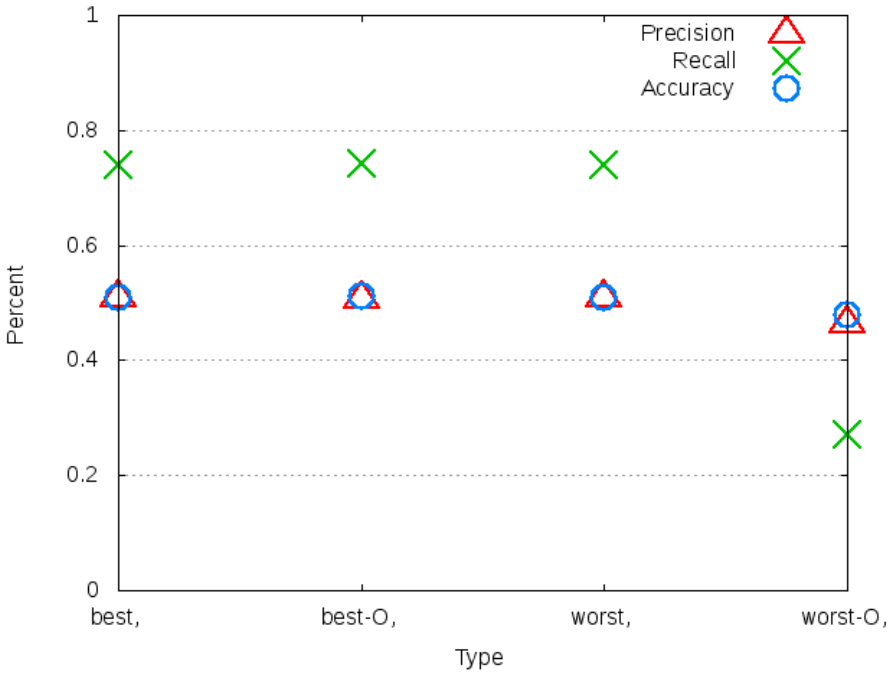


Figure 5.37: Oversampling for greenDAO using RF

5.3.2.4 Random Forest Discussion

The approach was experimented on using the machine learning algorithm RF. The three factors; SWR, feature set and OS were investigated. The first factor, SWR, achieved positive results for some of the projects and appeared to have the largest impact on the performance. While the impact of the SWR was larger often the best result for a project would of accuracy and precision at 0.5 and recall at 1.0. The impact of the feature set should also not be overshadowed as some projects performed better in the third experiment than the first. Finally, the impact of sample balancing on the data set proved primarily negative with both the best and worst trails performing worse when OS was used.

The first experimental results did not show a pattern for present for all the projects. Some projects performed better while a lot performed mediocre at just above 0.5 precision, accuracy. A few of the projects had poor performances with measures dropping below 0.5. Most of the groups yielded positive similar results, with acra, dagger and ShowcaseView all performing well and having fairly large variations for different trials. Likewise, aqruillian-core, brave and spark all performed similarly. While the results for these three project were not particularly good they still showed similar trends of performance.

For the second experiment no particular feature set was found to work for all projects. Some projects performed well for some while others performed well for others. Generally though, projects tended not perform well for all feature sets. One project such project to perform poorly was governator which performed well for precision, okay for accuracy and terribly for recall. However in terms of project groupings, the grouping of acra, dagger and ShowcaseView all performed well but varied in which feature set they performed well with. Overall groups tended to perform differently and had success with different feature sets. A feature set that worked well for one project

in the group could just as easily work poorly for another. This is highlighted best in the group consisting of http-request, nettosphere, parceler and retrolamdba. Both http-request and parceler perform well and nettosphere and retrolamdba perform for badly all feature sets.

Of course the final experiment with oversampling performed similar to the SVM experiments. For most projects the use of OS in balancing the sample data provided no difference in model performance. The second most common outcome was for trails that used OS to result in lower performance than without. Finally in some rare cases, 2 of the 23 projects saw a one of their trials perform slightly better with the use of OS. The results were not consistent for these projects since the first project was dagger saw the other corresponding trail decrease in performance. However the second project, smile, recorded no difference in result for the corresponding trail.

Overall for the experiments using RF, the variable with the greatest impact was the SWR. However these variables were not consistent across projects. Even for projects that performed well the variable values differed. Projects that worked tended to be repositories that were long, small in size, with a small to medium number of developers and a low to medium rate of commits. This outcome is similar to the experiments conducted with SVM.

5.3.3 Experiment Discussions

The best performance each project out of the experiments conducted are shown in Table 5.19. Therefore the best parameters are outlined for each project. The best result was calculated through multi-variable optimization. Since the performance measure consisted of three variables the best result would be one that maintained higher results for each. While not the most ideal, a weighted sum was taken of the precision, recall and accuracy outlined in Equation 5.9. The weight of precision (w_p)

and accuracy (w_a) are closely related both were assigned a weight of 0.5 while the weight for recall (w_r) was assigned 1.0. The higher the resultant summation of the performance measures the higher the ranking of the given performance. Therefore the trails that had the highest performance weighted sum were considered to have performed the best.

$$score_i = w_{pi} \times p_i + w_{ri} \times r_i + w_{ai} \times a_i max(score) \tag{5.9}$$

While the weighted sum did optimize each value, a common issue was that often a project would have one or two parameter perform well while the remaining performed poorly. For example blockly-android had the maximum value for recall but 0.51 for precision and accuracy. The weighted sum of this vector would be $0.51 \times 0.5 + 1.0 \times 1.0 + 0.51 \times 0.5 = 1.51$. However in the event that another performance scored 0.7 for each measure, the weighted sum would be $0.7 \times 0.5 + 0.7 \times 1.0 + 0.7 \times 0.5 = 1.4$. Since the goal is for multi-variable optimization a model that is generally good on each measure is better than a model performs well with one or two but terribly for the rest. Regardless though, some generally performed poorly and that is reflected in the table with a best performance that is quite low or very bias.

The number of projects that perform the best with SVM and RF are 11 and 12 respectively. As noted above some of the performance results are still quite low for some of the projects because of a lack of success for that project for all of the trials run. All projects did perform better than 50% which is quite a poor result since it is about as good as a coin toss. Out of the best performances, only 6 of the 23 projects have a trial that performs better than 60% for all three measures. Of those 6 only 4 of those perform better than 70% for each performance measure. Of the top 4, http-request performed the best followed by acra, ShowcaseView and ion in descending

Project	AI	Feature Set	SWR	Test Setup	Precision	Recall	Accuracy
acra	SVM	2	80	test 1	0.74	0.92	0.8
arquillian-core	RF	3	90	test 3	0.53	0.98	0.55
blockly-android	SVM	2	60	test 1	0.51	1.0	0.51
brave	RF	2	110	test 1	0.59	0.97	0.65
cardslib	SVM	2	120	test 1	0.5	1.0	0.5
dagger	RF	3	90	test 3	0.5	1.0	0.5
deeplearning4j	RF	2	70	test 1	0.55	0.96	0.58
fresco	RF	2	60	test 1	0.52	1.0	0.53
governator	RF	2	60	test 1	0.57	0.81	0.6
greenDAO	SVM	4	90	test 3	0.5	1.0	0.5
http-request	RF	2	80	test 1	0.87	0.79	0.84
ion	RF	2	90	test 3	0.72	0.73	0.72
jadx	SVM	2	130	test 1	0.55	0.82	0.58
mapstruct	SVM	2	70	test 1	0.6	0.88	0.65
nettosphere	RF	2	110	test 1	0.63	0.67	0.63
parceler	SVM	1	90	test 3	0.57	0.92	0.61
retrolambda	SVM	2	130	test 1	0.5	1.0	0.5
ShowcaseView	SVM	1	90	test 3	0.73	0.89	0.78
smile	SVM	2	70	test 1	0.5	1.0	0.5
spark	SVM	4	90	test 3	0.5	1.0	0.5
storm	RF	2	60	test 1	0.61	0.89	0.66
tempto	RF	2	120	test 1	0.53	0.73	0.55
yardstick	SVM	2	70	test 1	0.55	0.79	0.57

Table 5.19: Project Best Performance

order. Both acra and ShowcaseView performed best with use of SVM while http-request and ion perform best with RF. For both http-request and ShowcaseView the overall performance is better for RF over SVM. Likewise, ion performed better for SVM over RF. The results were more consistent between RF and SVM. The second feature set used for acra, http-request and ion to gain the best performance while for ShowcaseView the first feature set performed the best. On a final note, each of these 4 repositories are classified as long in length, small in size, small to medium in number of developers and low to medium in commit rate.

5.4 Threats to Validity

This wider experimentation also proved to be very beneficial for the analysis of the method since performance was not consistent across all projects. A concerted effort was made to contrasting positive results for one project with negative results. Such a contrast may mitigate the impact of the positive results, however provide the full context and help direct future work in this area.

Each experiment was designed to attempt to provide a robust setup to measure accurately the performance of the approach given the changes to the current factor. The setup was designed to attempt to preventing the influence of other variables beyond the independent variable. The factors that may have had an influence on the experimental results are the third experiment only sampling the extremes (best and worst).

A major concern with the final experiment was that of the sampling of the best and worst results from the previous experiments to test the use of OS. While the results of the use of OS should not be discounted, only the sampling the extremes of the previous experiment may have limited the measurable impact of the use of OS. This experiment could be extended to test the middle performance or even test each trail from the previous experiments.

The differences between the projects prevented a more direct comparison between the projects. Furthermore, as shown through the experiments, some projects (e.g. ShowcaseView) generally performed better than other projects (e.g. governator). This leads to the conclusion that certain project related factors have a large impact on the performance of the approach. Further investigation into these project specific factors could lead to improved results for the approach.

Chapter 6

Conclusions

We proposed a method that leverages the commit history to predict future changes within the project. The changes that are predicted are in the short term of 5 commits. The approach was then tested on 23 different OSS projects developed in Java. The tests investigated the different factors that impacted the performance of the approach. The results of the tests show that while the SWR had a strong impact on the performance, the projects themselves often had internal factors which caused differences in performance.

The contributions of this work are:

1. Providing an approach that with some success can predict future changes within a project using the commit data. Both SVM and RF are viable for some projects.

2. Determined which factors more strongly influence the performance of the predictions. Out of the three factors investigated, the SWR proved to have the greatest impact for both SVM and RF.

Future work includes investigating the projects and they differ in the way they

change. The four projects that tended to do will together were acra, ShowcaseView, http-request and ion which also shared similar project characteristics. Finally a more extensive look at the other factors that were involved in the approach to determine their impact of the approach.

Bibliography

[1] Alam, M. S., and Vuong, S. T. Random Forest Classification for Detecting Android Malware. In *Proceedings of the Green Computing and Communications, IEEE Internet of Things, IEEE Cyber, Physical and Social Computing* (2013), pp. 663–669.

[2] Antón, J. C. Á., Nieto, P. J. G., Viejo, C. B., and Vilán, J. A. V. Support Vector Machines Used to Estimate the Battery State of Charge. *IEEE Transactions on Power Electronics 28*, 12 (2013), 5919 – 5926.

[3] Bantelay, F., Zanjani, M. B., and Kagdi, H. Comparing and combining evolutionary couplings from interactions and commits. In *Proceedings of the Working Conference on Reverse Engineering, WCRE* (2013), pp. 311–320.

[4] Bhattacharyya, S., Jha, S., Tharakunnel, K., and Westland, J. C. Data mining for credit card fraud: A comparative study. *Decision Support Systems 50*, 3 (2011), 601–613.

[5] Bieman, J., Andrews, A., and Yang, H. Understanding change-proneness in OO software through visualization. In *Proceedings of the 11th IEEE International Workshop on Program Comprehension, 2003.* (2003), pp. 44 – 53.

[6] Burbidge, R., Trotter, M., Buxton, B., and Holden, S. Drug design by machine learning : support vector machines for pharmaceutical data analysis. *Computers and Chemistry 26*, 1 (2001), 5 – 14.

[7] Chaturvedi, K. K., Kapur, P. K., Anand, S., and Singh, V. B. Predicting the complexity of code changes using entropy based measures. *International Journal of System Assurance Engineering and Management 5*, 2 (2014), 155–164.

[8] Collberg, C., Kobourov, S., Nagra, J., Pitts, J., and Wampler, K. A system for graph-based visualization of the evolution of software. In *Proceedings of the 2003 ACM symposium on Software visualization - SoftVis '03* (2003), pp. 77 – 86.

[9] De Souza, C. R., Quirk, S., Trainer, E., and Redmiles, D. F. Supporting collaborative software development through the visualization of socio-technical dependencies. *2007 International ACM Conference on Supporting Group Work, GROUP'07, November 4, 2007 - November 7, 2007* (2007), 147–156.

[10] Dit, B., Holtzhauer, A., Poshyvanyk, D., and Kagdi, H. A dataset from change history to support evaluation of software maintenance tasks. In *Proceedings of the 10th Working Conference on Mining Software Repositories* (2013), pp. 131–134.

[11] Erturk, E., and Akcapinar, E. A comparison of some soft computing methods for software fault prediction. *Expert Systems with Applications 42*, 4 (2015), 1872–1879.

[12] Gall, H. C., and Lanza, M. Software Evolution : Analysis and Visualization. In *Proceedings of the 28th international conference on Software engineering* (2006), pp. 1055–1056.

[13] Giger, E., Pinzger, M., and Gall, H. C. Can we predict types of code changes? An empirical analysis. In *Proceedings of the 9th IEEE Working Conference on Mining Software Repositories (MSR), 2012* (2012), pp. 217–226.

[14] Gilbert, E., and Karahalios, K. CodeSaw: A social visualization of distributed software development. In *Proceedings of the 11th IFIP TC 13 international conference on Humancomputer interactionVolume Part II* (2007), pp. 303–316.

[15] Gondra, I. Applying machine learning to software fault-proneness prediction. *Journal of Systems and Software 81*, 2 (2008), 186–195.

[16] Gonzalez, A., Theron, R., Telea, A., and Garcia, F. J. Combined Visualization of Structural and Metric Information for Software Evolution Analysis. In *Proceedings of the joint international and annual ERCIM workshops on Principles of software evolution (IWPSE) and software evolution (Evol) workshops* (2009), pp. 25–29.

[17] Granitto, P. M., Gasperi, F., Biasioli, F., and Furlanello, C. Modern data mining tools in descriptive sensory analysis: A case study with a Random forest approach. *Food Quality and Preference 18*, 4 (2007), 681–689.

[18] Guo, L., Ma, Y., Cukic, B., and Singh, H. Robust Prediction of Fault-Proneness by Random Forests. In *Proceedings of the 15th International Symposium on Software Reliability Engineering, 2004* (2004), pp. 417–428.

[19] Hassan, A. E. Mining software repositories to assist developers and support managers. In *Proceedings of the 22nd IEEE International Conference on Software Maintenance* (2006), pp. 339–342.

[20] Hassan, A. E., and Holt, R. C. Predicting change propagation in software systems. In *20th IEEE International Conference on Software Maintenance, 2004* (2004), pp. 284–293.

[21] Hemmati, H., Nadi, S., Baysal, O., Kononenko, O., Wang, W., Holmes, R., and Godfrey, M. W. The MSR cookbook: Mining a decade of research. In *Proceedings of the 10th Working Conference on Mining Software Repositories* (2013), pp. 343–352.

[22] Huang, C.-l., Chen, M.-c., and Wang, C.-j. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications 33*, 4 (2007), 847–856.

[23] Jalbert, K., and Bradbury, J. S. Predicting mutation score using source code and test suite metrics. In *Proceedings of the 1st International Workshop on Realizing AI Synergies in Software Engineering* (jun 2012), Ieee, pp. 42–46.

[24] Kagdi, H., and Maletic, J. I. Combining single-version and evolutionary dependencies for software-change prediction. In *Proceedings of the 4th International Workshop on Mining Software Repositories, MSR 2007* (2007).

[25] Keim, D. a. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics 8*, 1 (2002), 1–8.

[26] Khoshgoftaar, T. M., Golawala, M., and Van Hulse, J. An Empirical Study of Learning from Imbalanced Data Using Random Forest. In *Proceedings*

of the 19th IEEE International Conference on Tools with Artificial Intelligence (2007), pp. 310–317.

[27] Kim, K.-j. Financial time series forecasting using support vector machines. *Neurocomputing 55*, 1-2 (2003), 307–319.

[28] Kim, S., Jr, E. J. W., and Zhang, Y. Classifying Software Changes : Clean or Buggy ? *IEEE Transactions on Software Engineering 34*, 2 (2008), 181–197.

[29] Maletic, J. I., and Collard, M. L. Supporting source code difference analysis. In *Proceedings of the 20th IEEE International Conference on Software Maintenance* (2004), pp. 210–219.

[30] Malhotra, R. A systematic review of machine learning techniques for software fault prediction. *Applied Soft Computing 27*, C (2015), 504–518.

[31] Moeyersoms, J., Fortuny, E. J. D., Dejaeger, K., Baesens, B., and Martens, D. Comprehensible software fault and effort prediction : A data mining approach. *Journal of Systems & Software 100* (2015), 80–90.

[32] Moser, R., Pedrycz, W., and Succi, G. A comparative analysis of the efficiency of change metrics and static code attributes for defect prediction. In *Proceedings of the 30th international conference on Software engineering* (2008), pp. 181–190.

[33] Murphy, C., Kaiser, G., and Arias, M. An Approach to Software Testing of Machine Learning Applications. In *Proceedings of the 19th International Conference on Software Engineering & Knowledge Engineering* (2007), pp. 167 – 172.

[34] Nagappan, N., and Ball, T. Using software dependencies and churn metrics to predict field failures: An empirical case study. In *Proceedings of the 1st International Symposium on Empirical Software Engineering and Measurement, 2007.* (2007), pp. 364–373.

[35] Neuhaus, S., Zimmermann, T., Holler, C., and Zeller, A. Predicting Vulnerable Software Components. In *Proceedings of the 14th ACM conference on Computer and communications security* (2007), pp. 529–540.

[36] Ogawa, M., and Ma, K.-L. StarGate: A Unified, Interactive Visualization of Software Projects. *2008 IEEE Pacific Visualization Symposium 3*, 11 (2008), 191 – 198.

[37] Ogawa, M., and Ma, K.-L. Software Evolution Storylines. In *Proceedings of the 5th international symposium on Software visualization* (2010), pp. 35 – 42.

[38] Sisman, B., and Kak, A. C. Incorporating version histories in Information Retrieval based bug localization. In *Proceedings of the 9th IEEE Working Conference on Mining Software Repositories* (jun 2012), Ieee, pp. 50–59.

[39] Thwin, M. M. T., and Quah, T.-S. Application of neural networks for software quality prediction using object-oriented metrics. *Journal of Systems and Software 76*, 2 (2005), 147–156.

[40] Verikas, A., Gelzinis, A., and Bacauskiene, M. Mining data with random forests: A survey and results of new tests. *Pattern Recognition 44*, 2 (2011), 330–349.

[41] Ying, A. T. T., Murphy, G. C., Ng, R., and Chu-Carroll, M. C. Predicting Source Code Changes by Mining Change History. *IEEE Transactions on Software Engineering 30*, 9 (2004), 574–586.

[42] Yu, G., Yaun, J., and Liu, Z. Unsupervised random forest indexing for fast action search. In *Computer Vision and Pattern Recognition* (2011), pp. 865 – 872.

[43] Zeng, J., and Qiao, W. Short-term solar power prediction using a support vector machine. *Renewable Energy 52* (2016), 118–127.

[44] Zimmermann, T., Weissgerber, P., Diehl, S., and Zeller, A. Mining Version Histories to Guide Software Changes. *IEEE Transactions on Software Engineering 31*, 6 (2005), 429–445.

Appendix A

Experimental Data

A.1 Experiment 1

A.1.1 Support Vector Machine

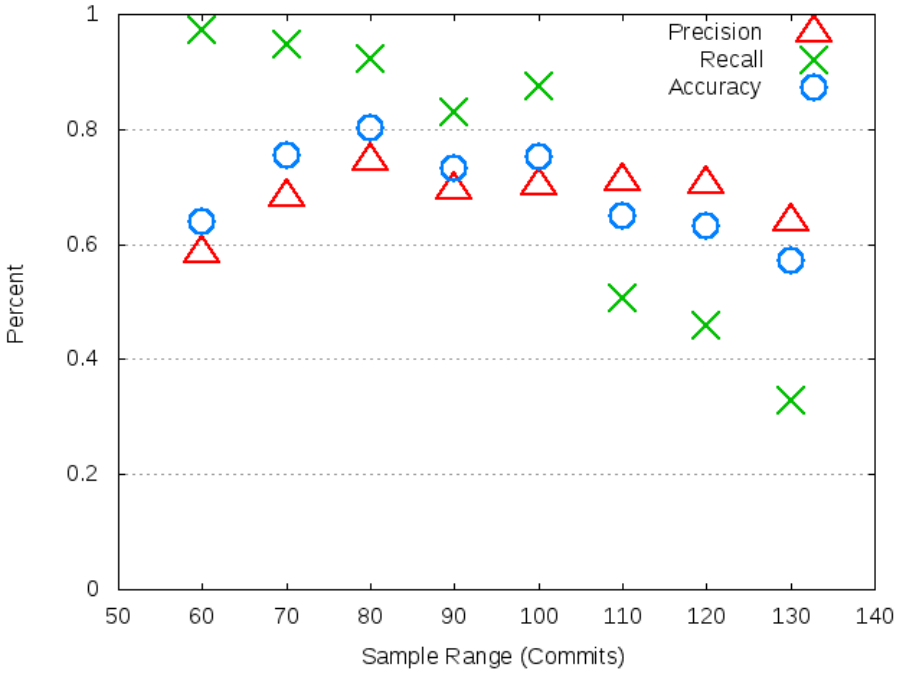


Figure A.1: SWR for acra using SVM

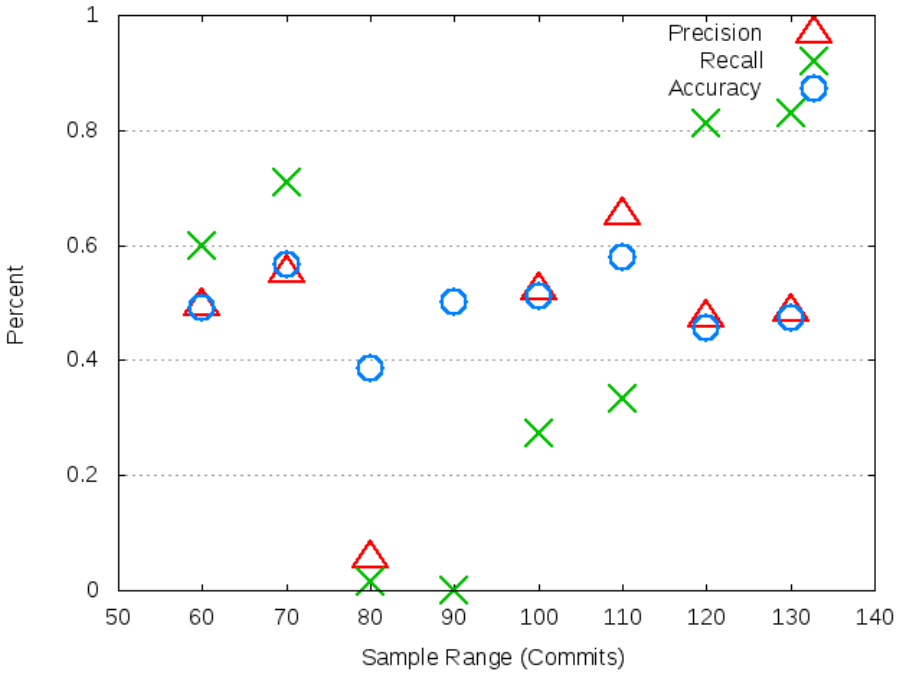


Figure A.2: SWR for arquillian-core using SVM

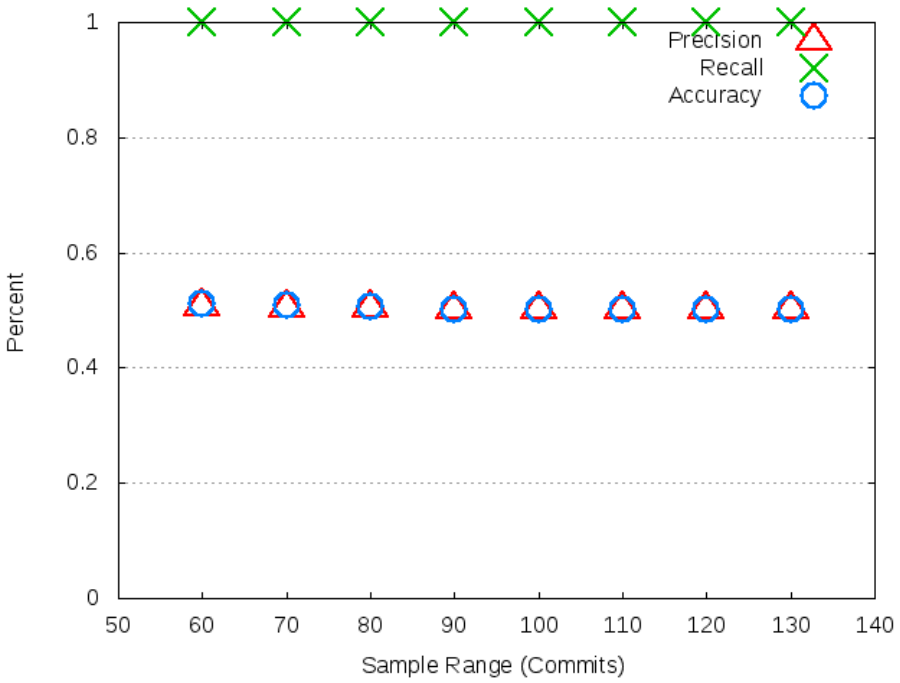


Figure A.3: SWR for blockly-android using SVM

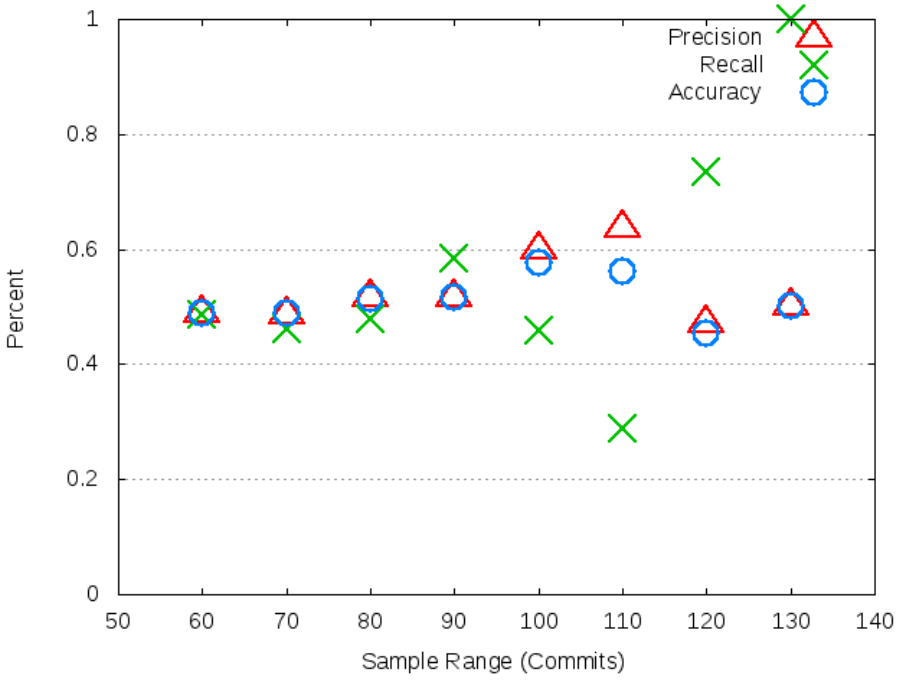


Figure A.4: SWR for brave using SVM

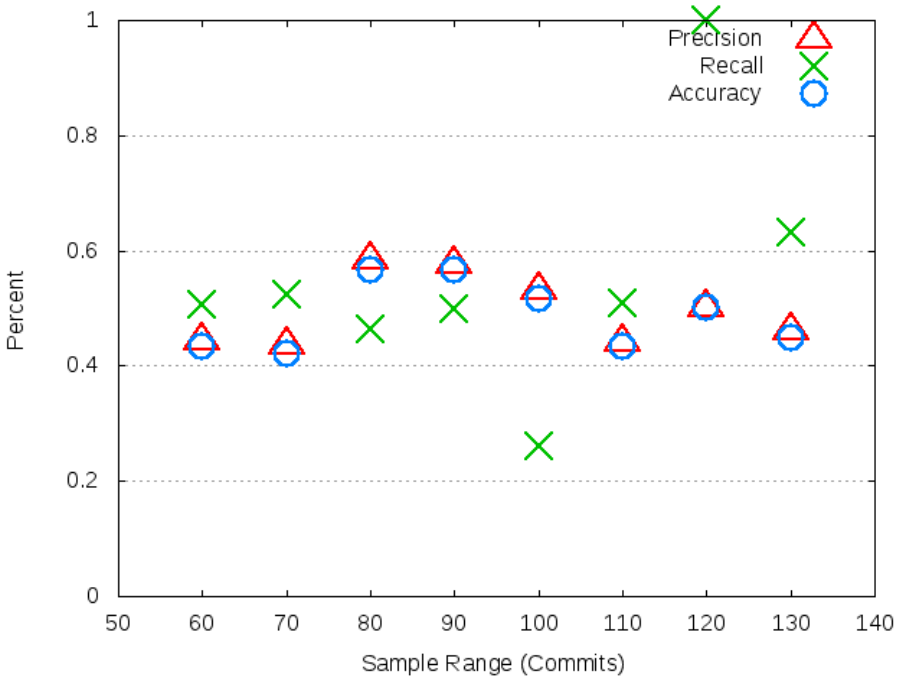


Figure A.5: SWR for cardslib using SVM

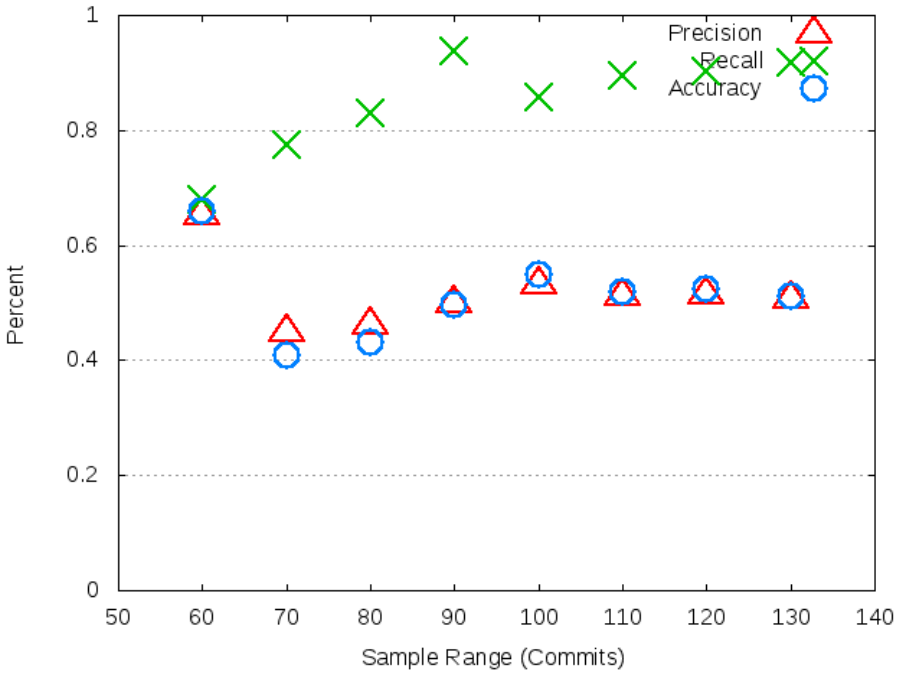


Figure A.6: SWR for dagger using SVM

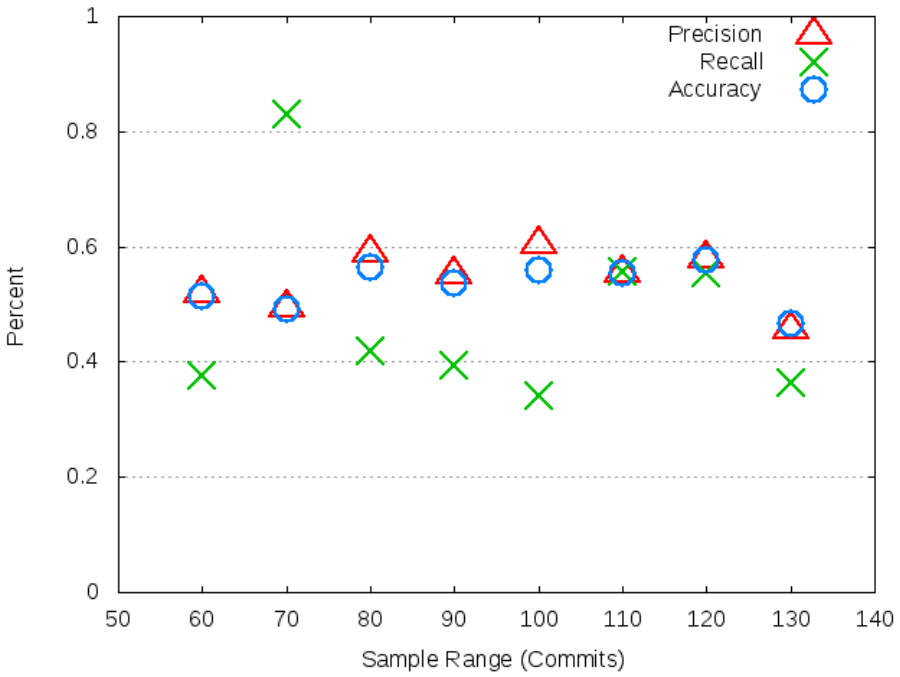


Figure A.7: SWR for deeplearning4j using SVM

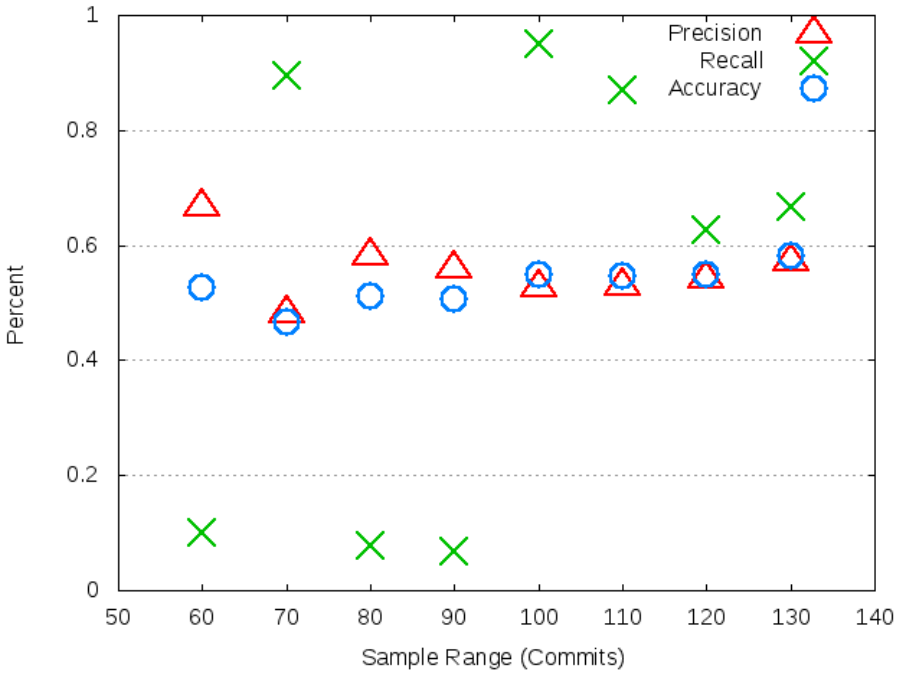


Figure A.8: SWR for fresco using SVM

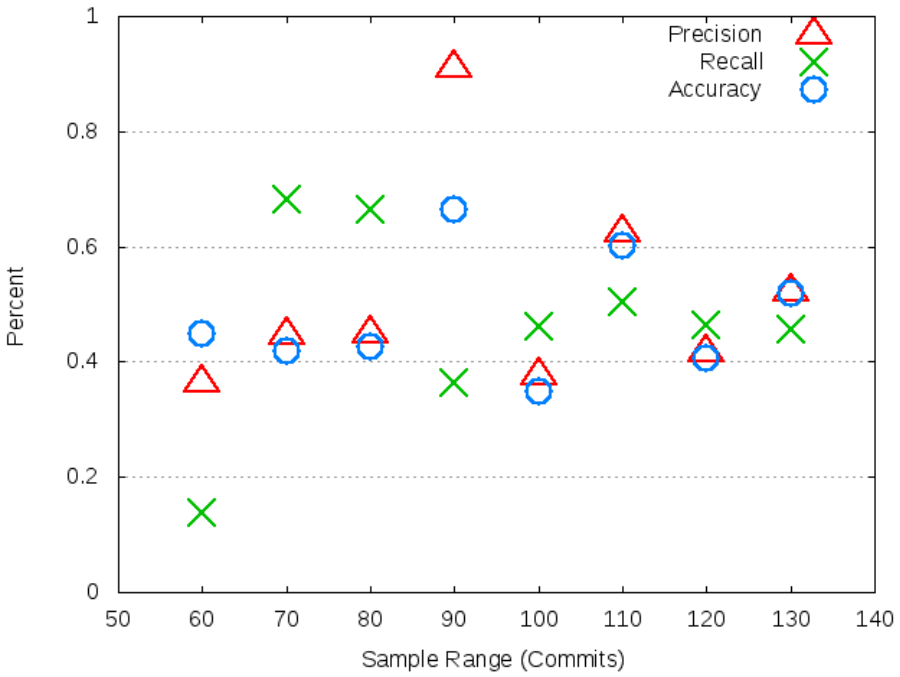


Figure A.9: SWR for governator using SVM

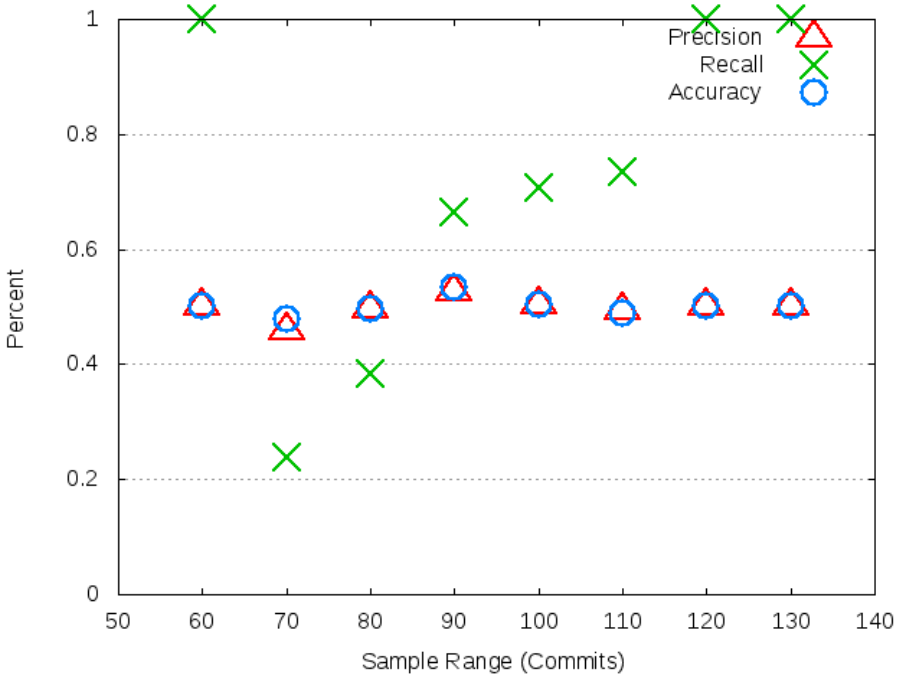


Figure A.10: SWR for greenDAO using SVM

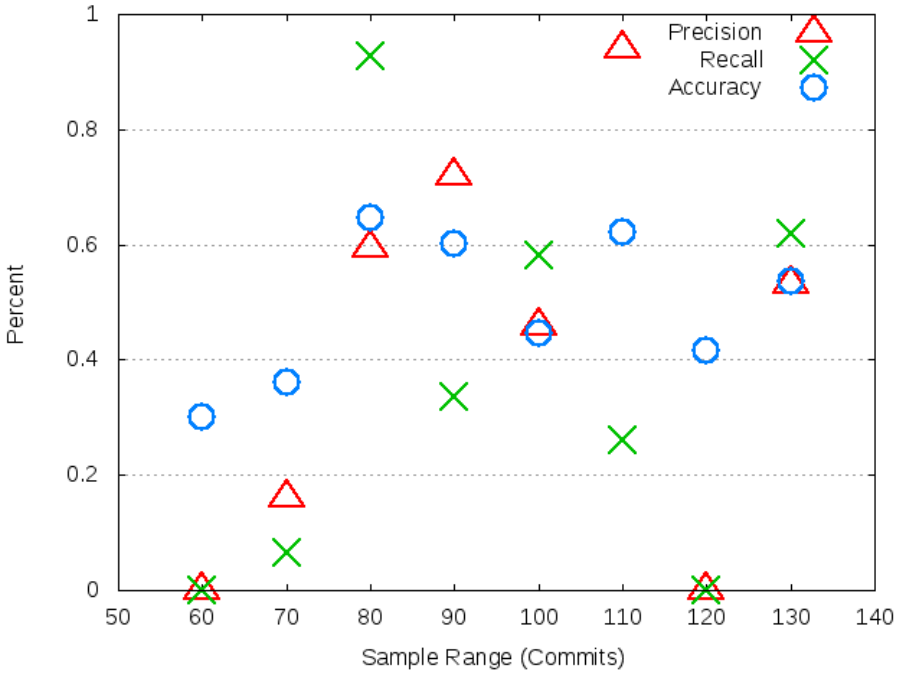


Figure A.11: SWR for http-request using SVM

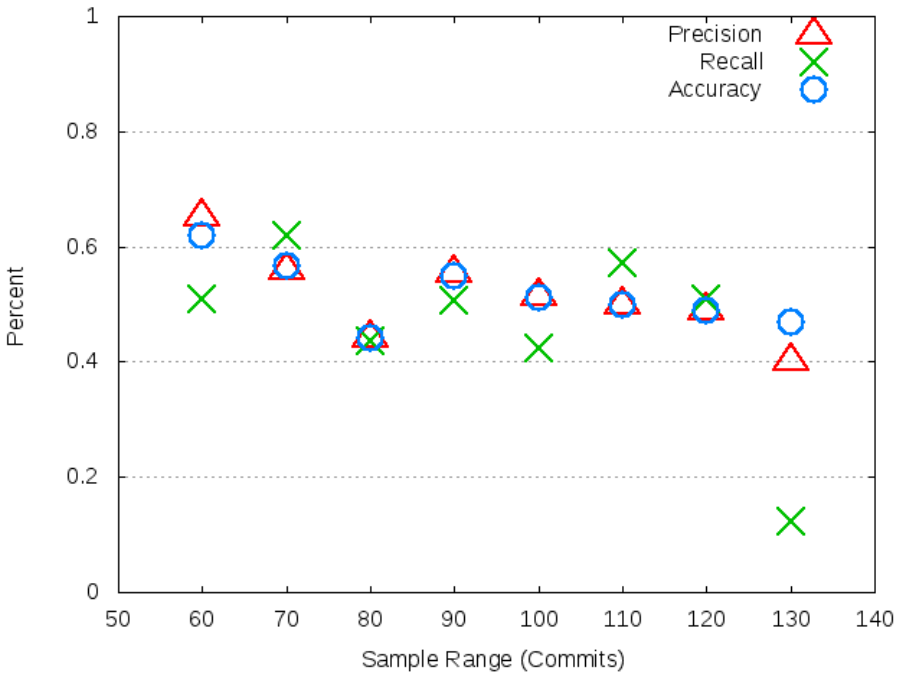


Figure A.12: SWR for ion using SVM

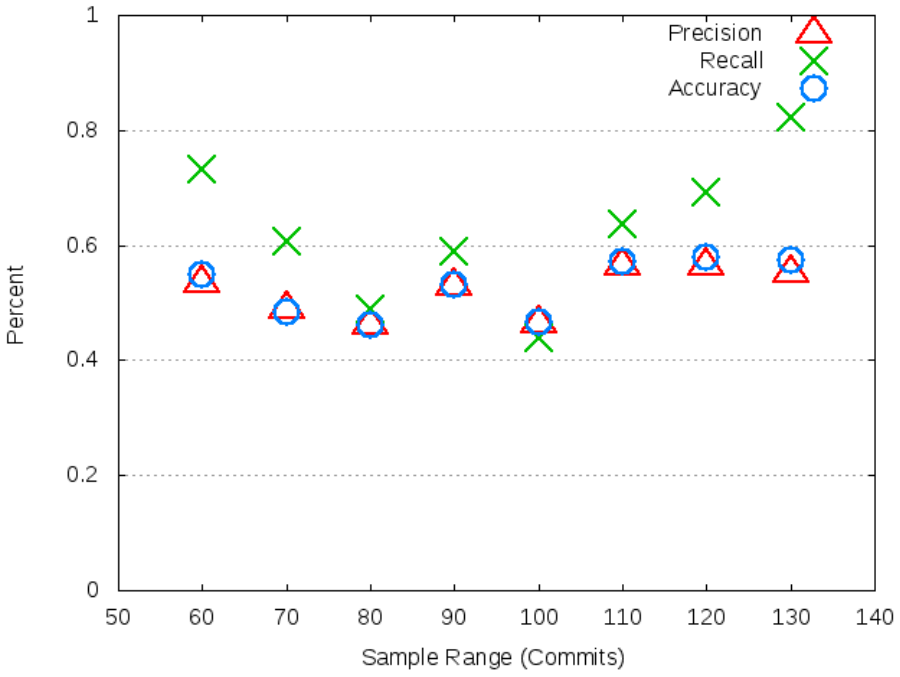


Figure A.13: SWR for jadx using SVM

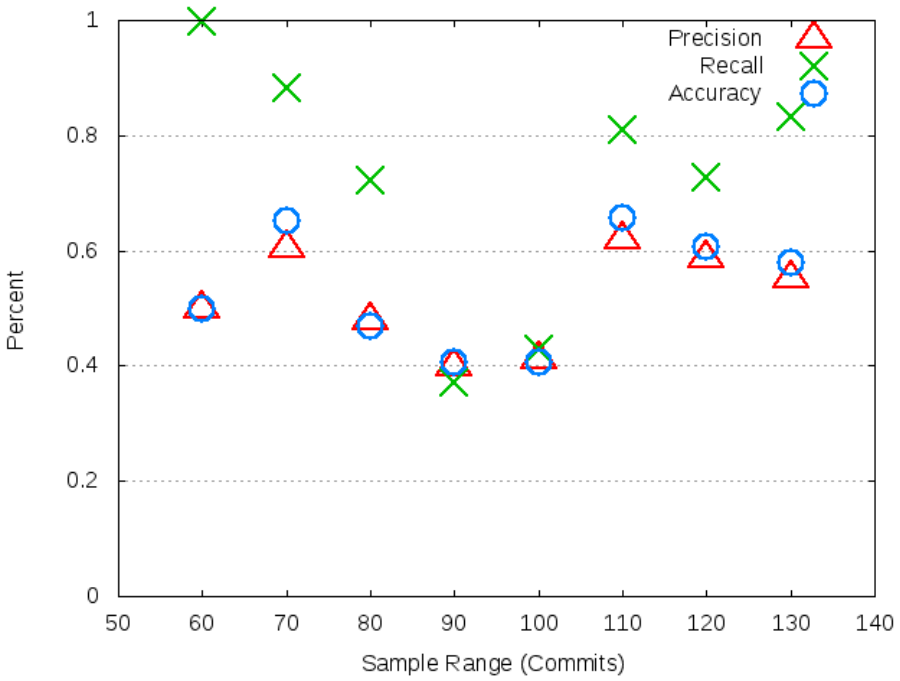


Figure A.14: SWR for mapstruct using SVM

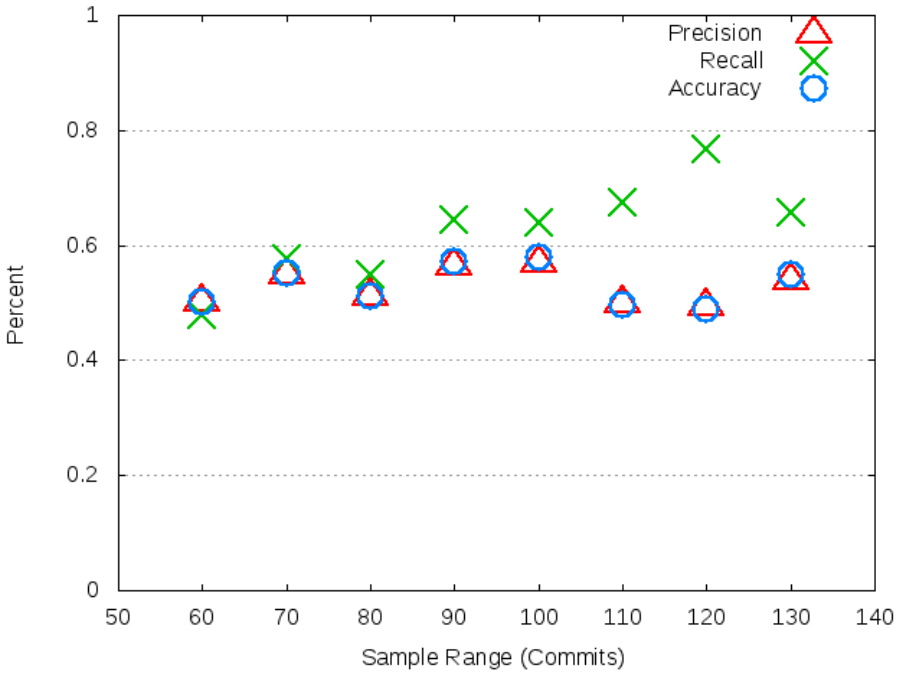


Figure A.15: SWR for nettosphere using SVM

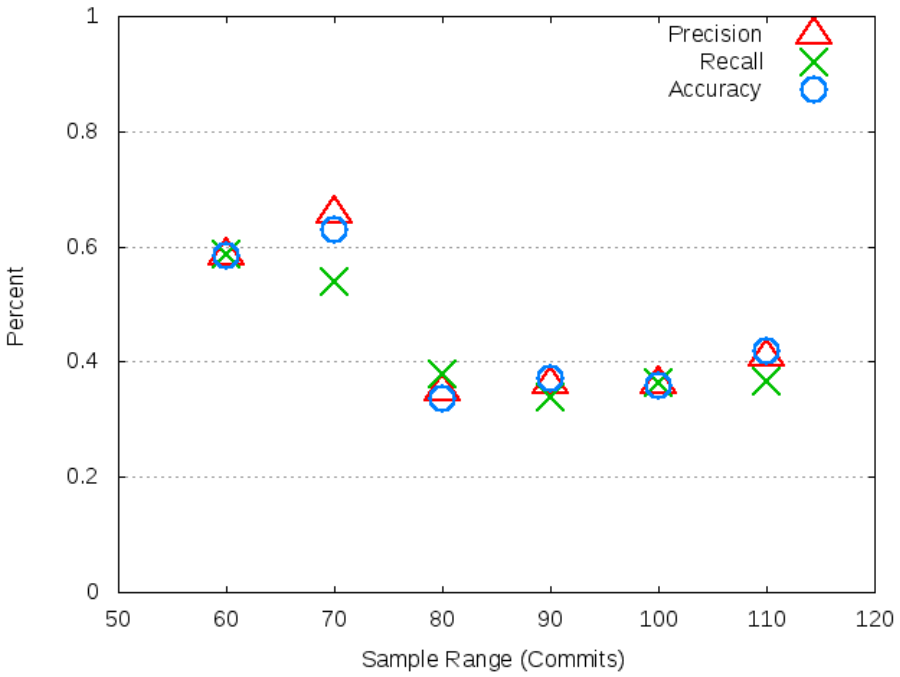


Figure A.16: SWR for parceler using SVM

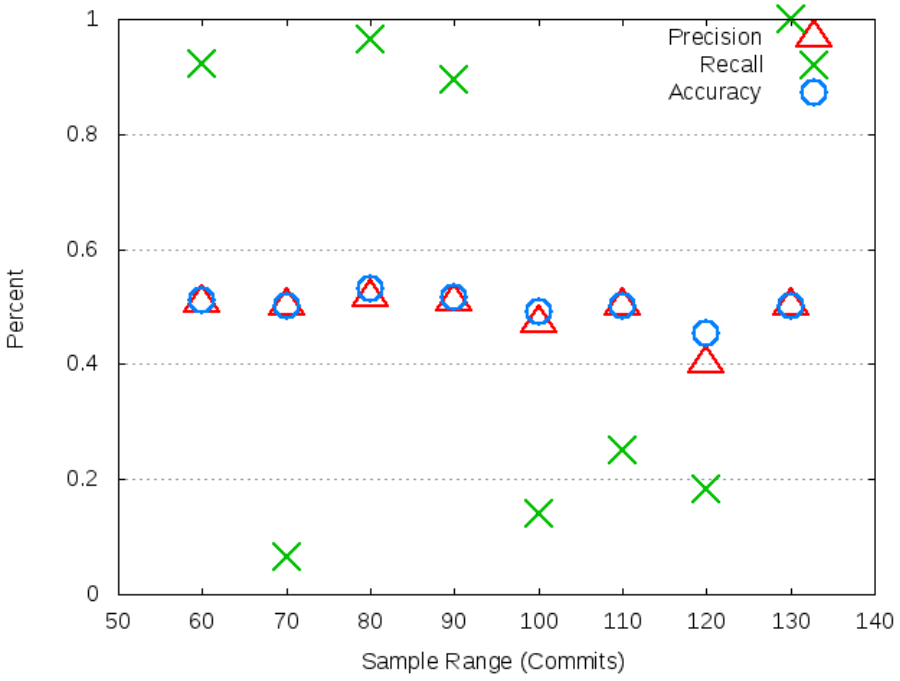


Figure A.17: SWR for retrolambda using SVM

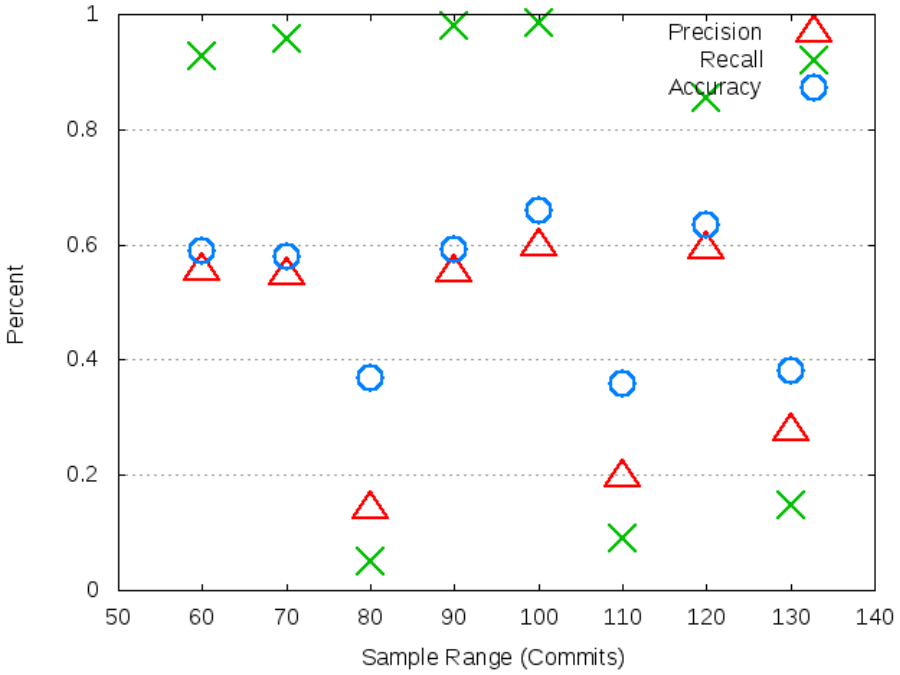


Figure A.18: SWR for ShowcaseView using SVM

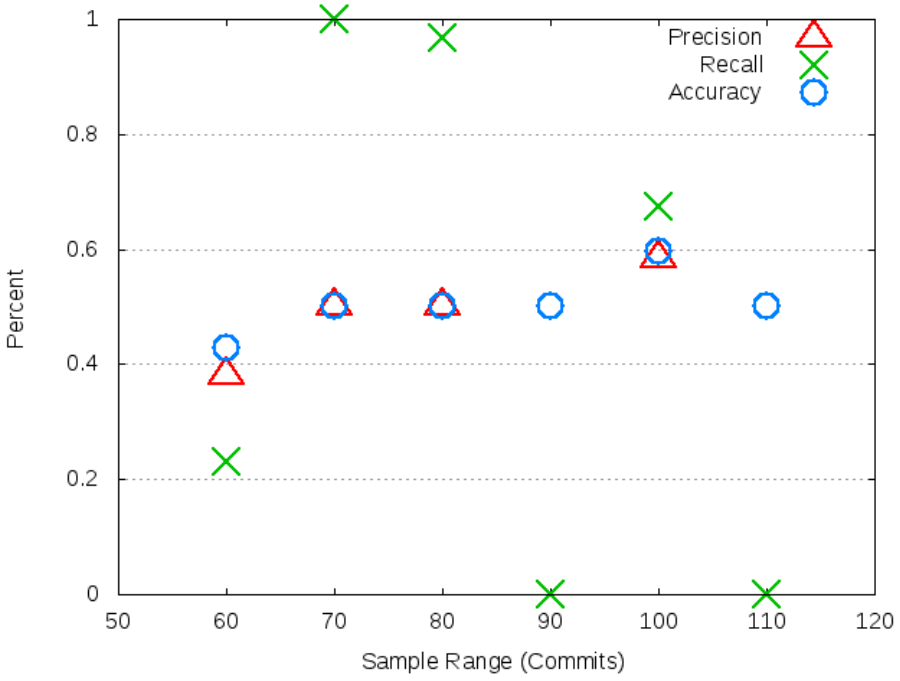


Figure A.19: SWR for smile using SVM

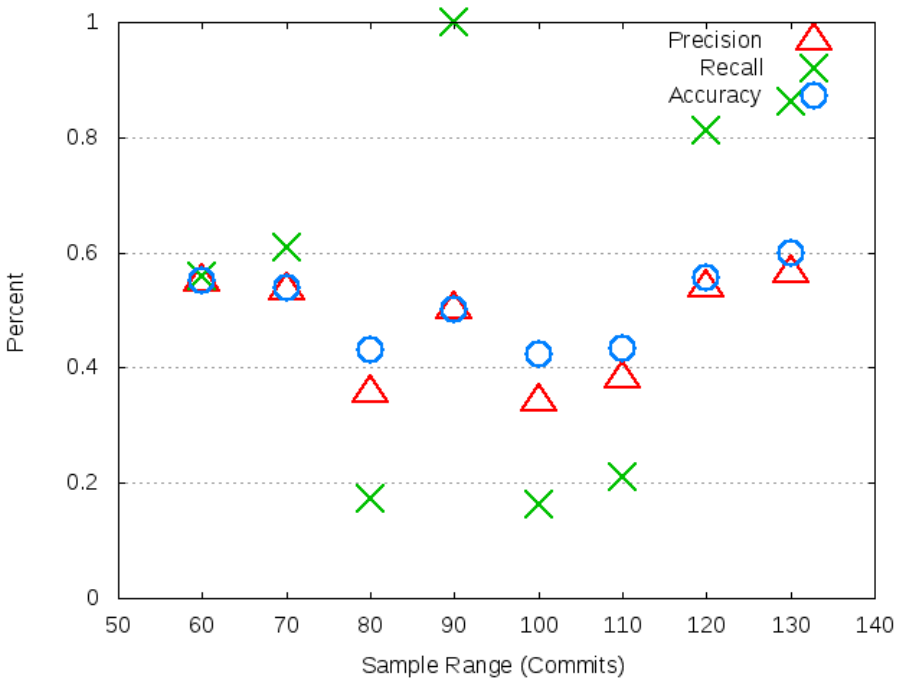


Figure A.20: SWR for spark using SVM

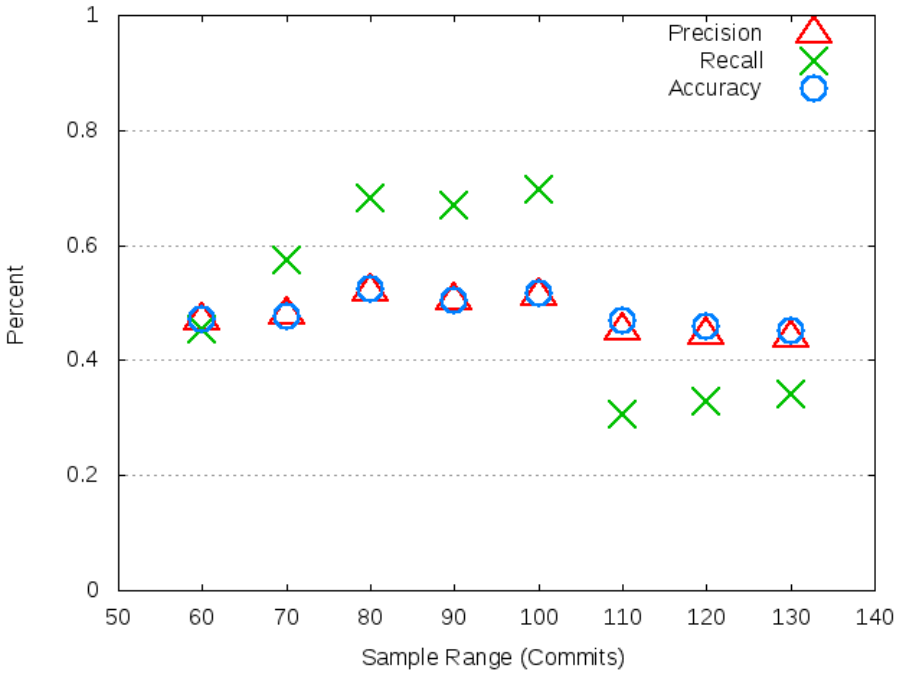


Figure A.21: SWR for storm using SVM

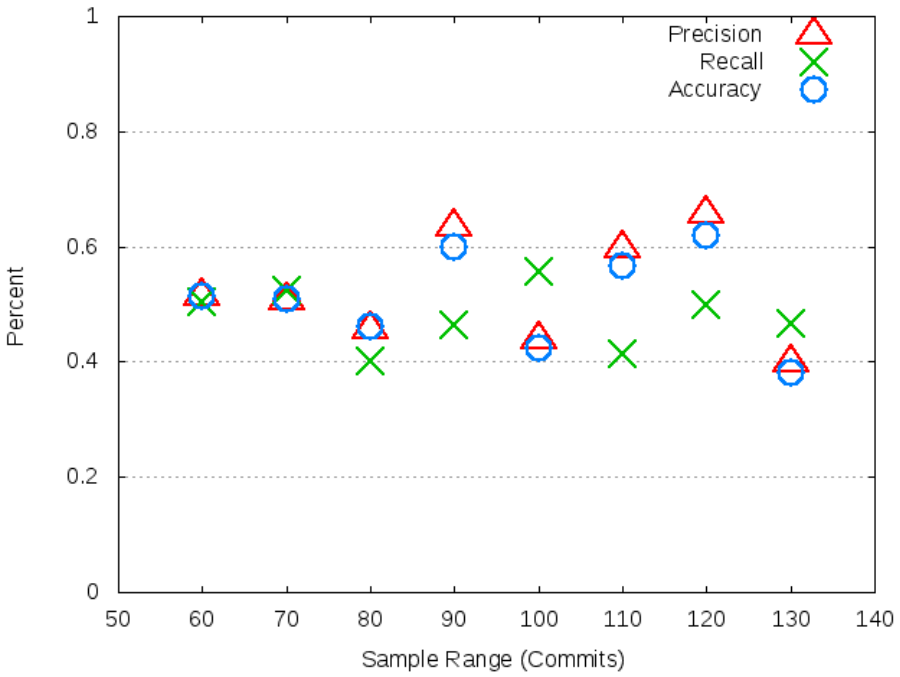


Figure A.22: SWR for tempto using SVM

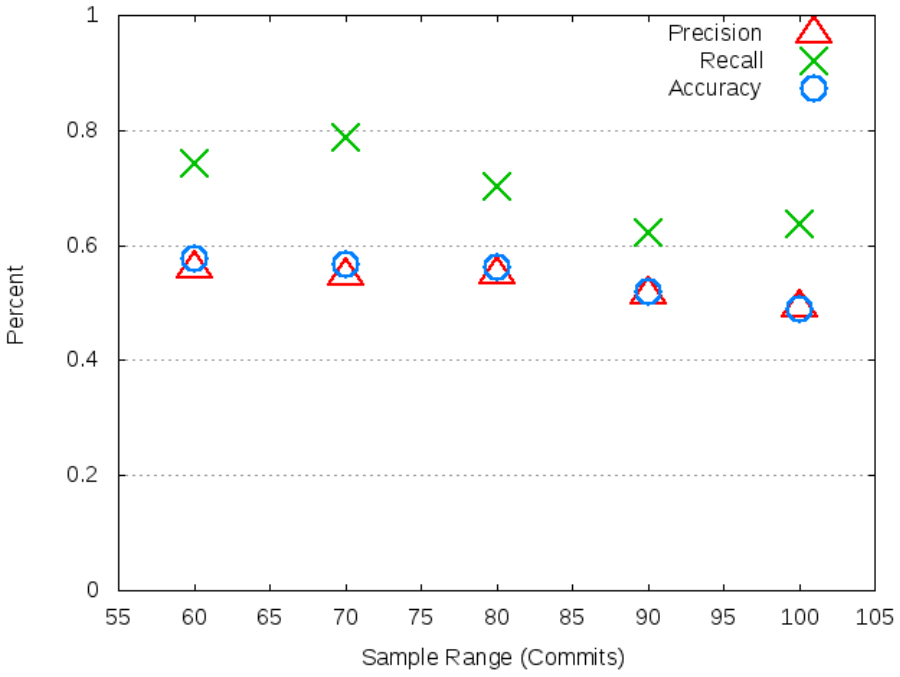


Figure A.23: SWR for yardstick using SVM

A.1.2 Random Forest

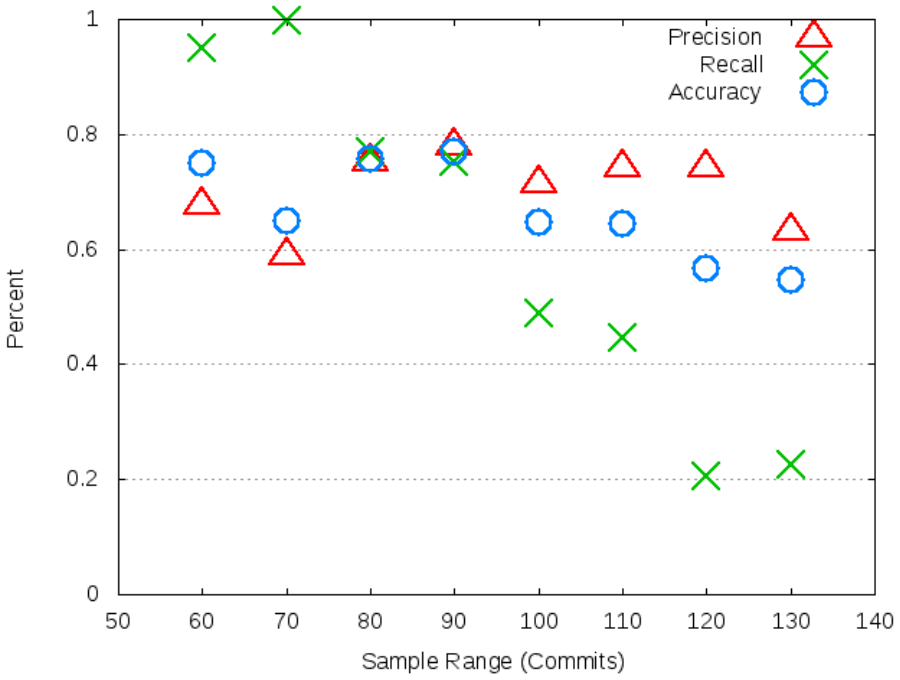


Figure A.24: SWR for acra using RF

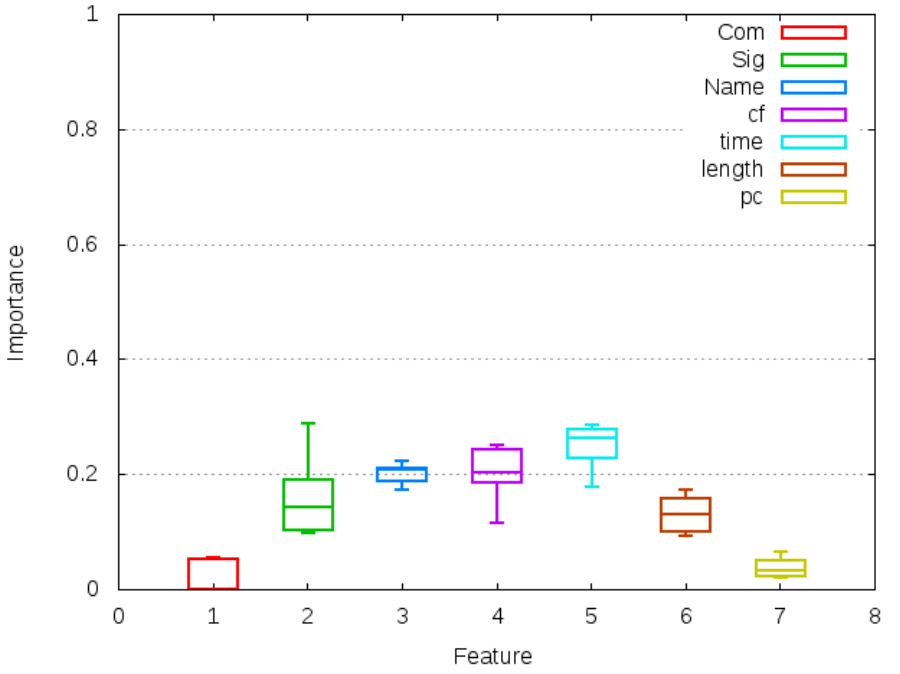


Figure A.25: Feature Importance SWR for acra using RF

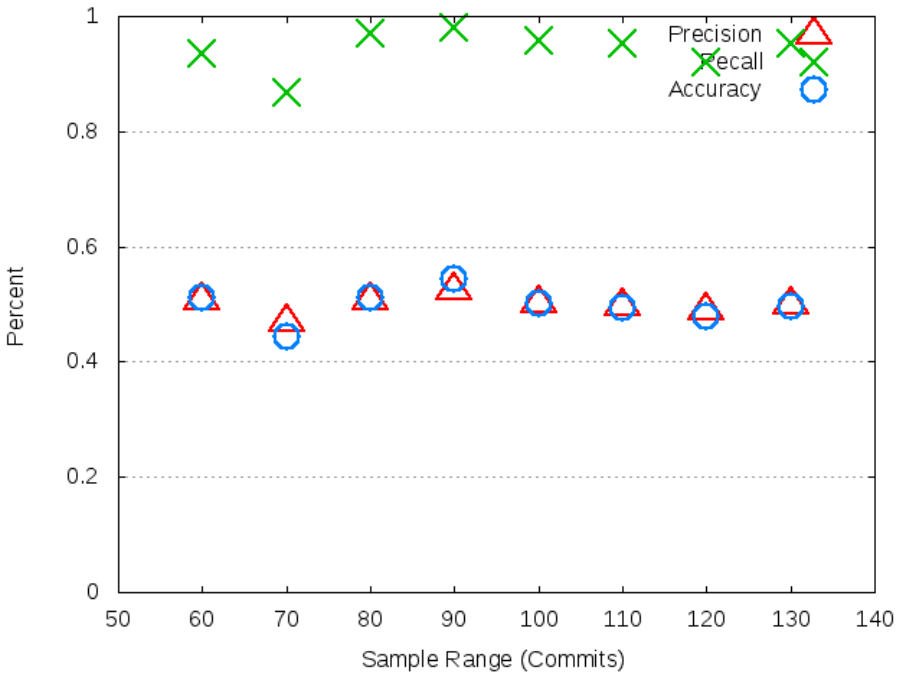


Figure A.26: SWR for arquillian-core using RF

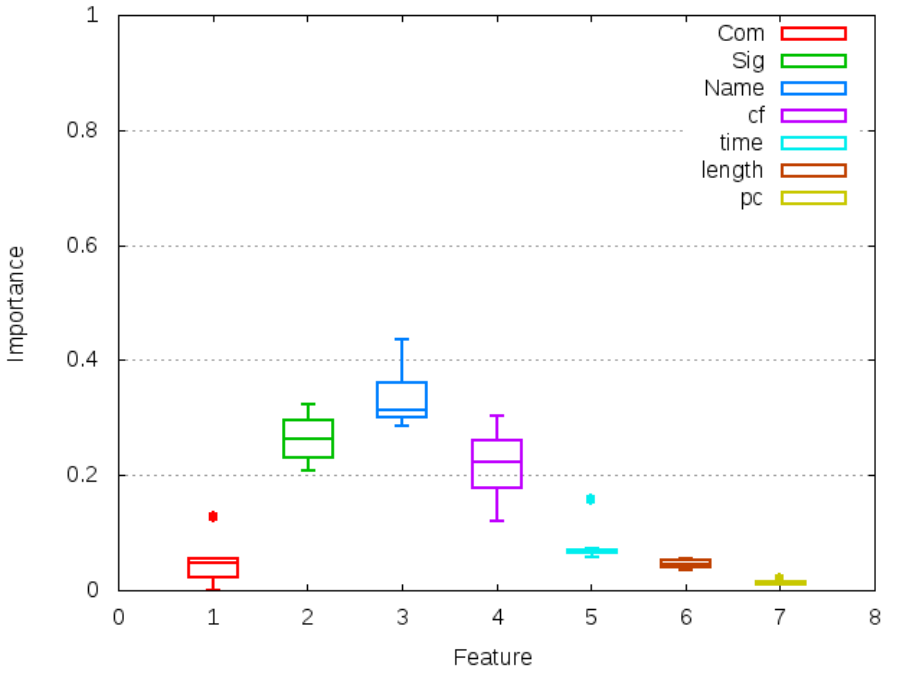


Figure A.27: Feature Importance SWR for arquillian-core using RF

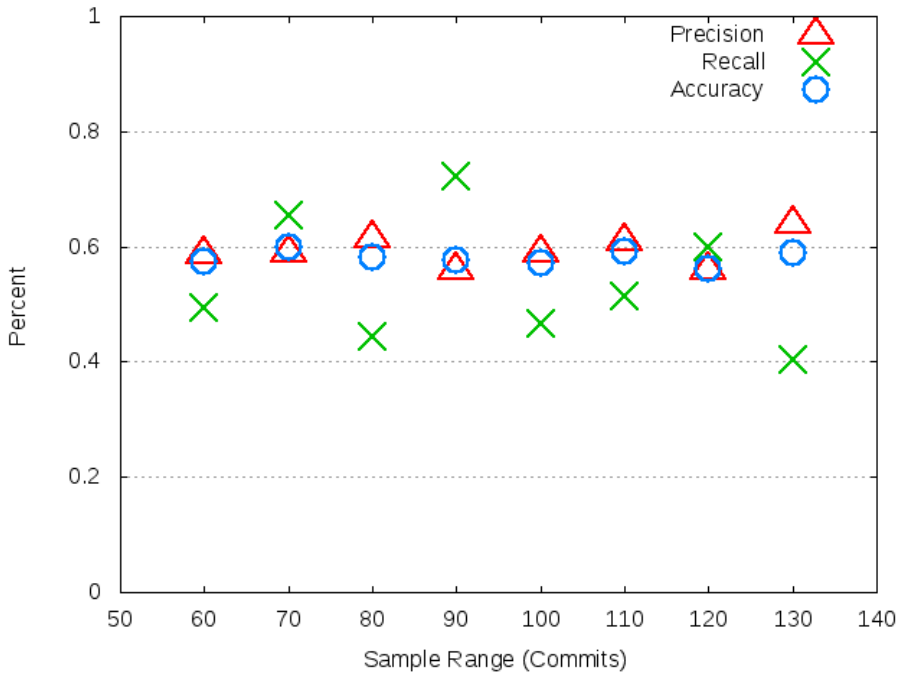


Figure A.28: SWR for blockly-android using RF

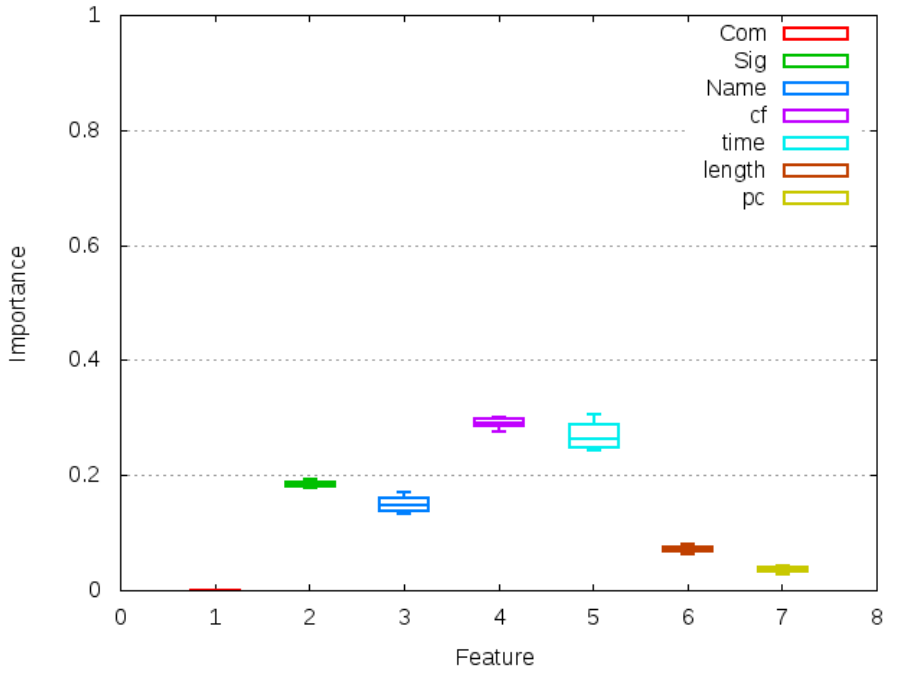


Figure A.29: Feature Importance SWR for blockly-android using RF

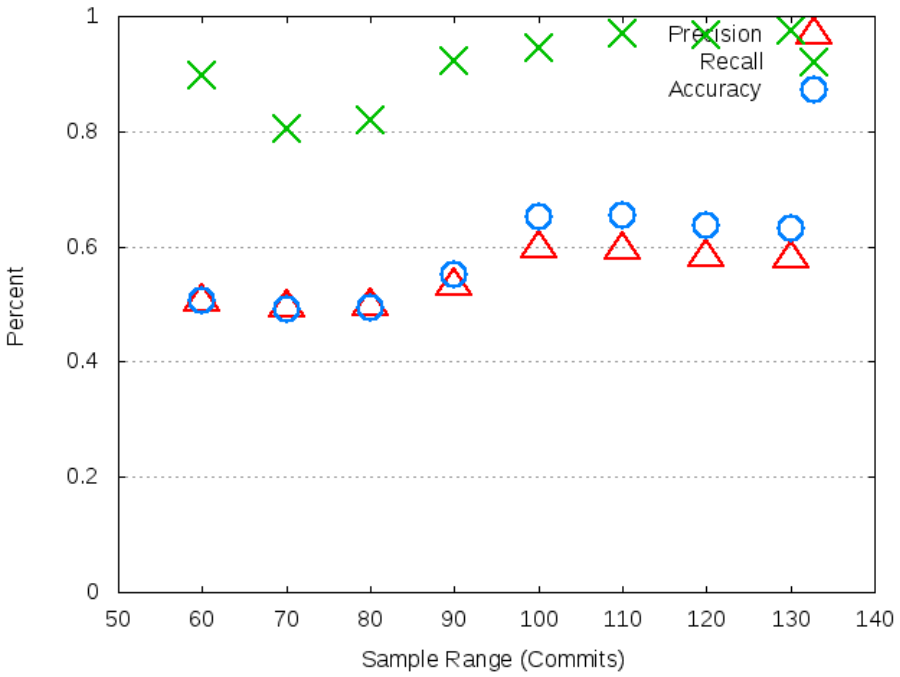


Figure A.30: SWR for brave using RF

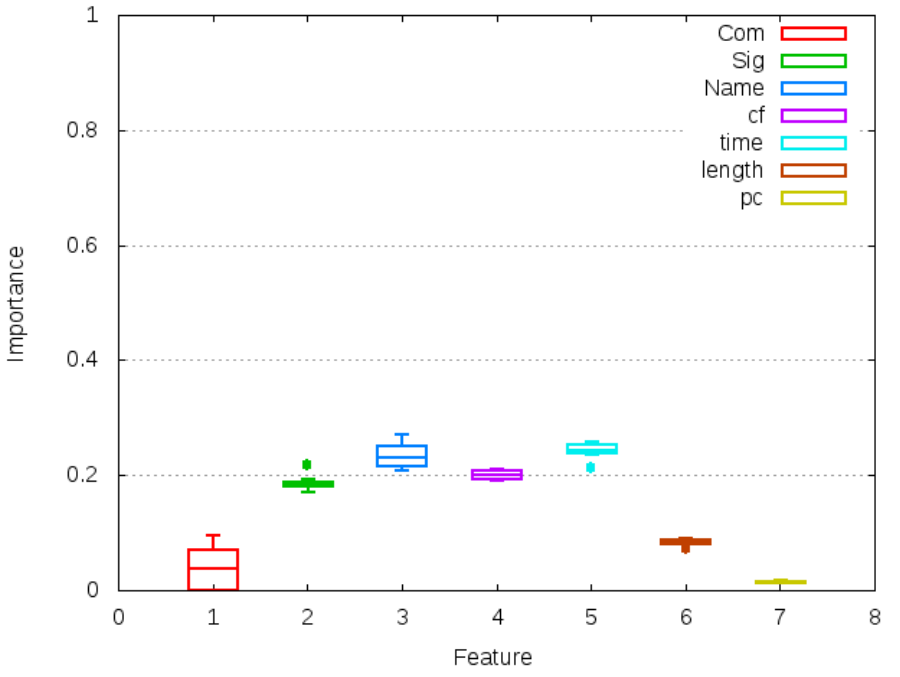


Figure A.31: Feature Importance SWR for brave using RF

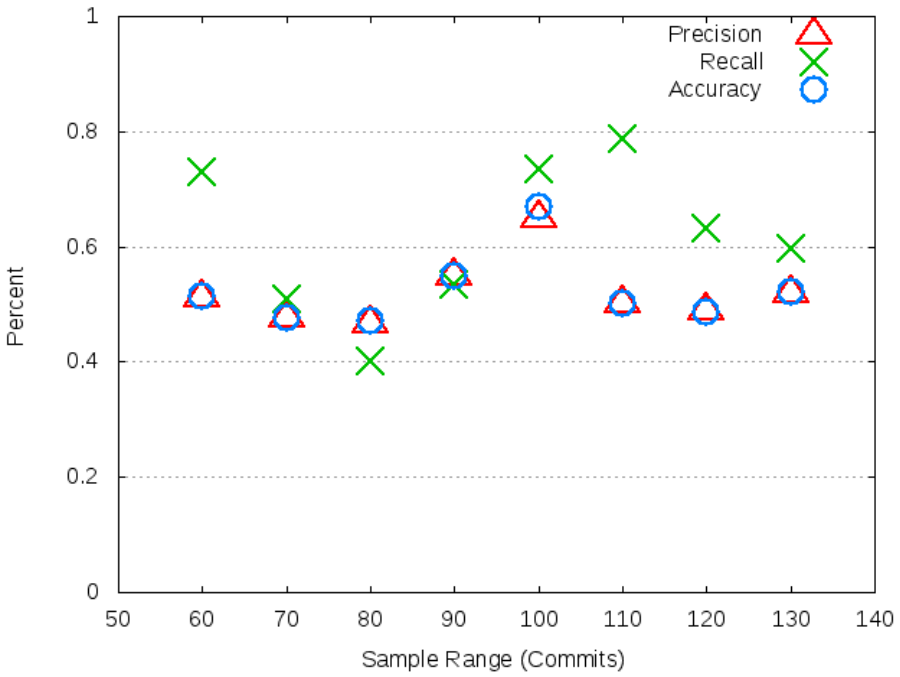


Figure A.32: SWR for cardslib using RF

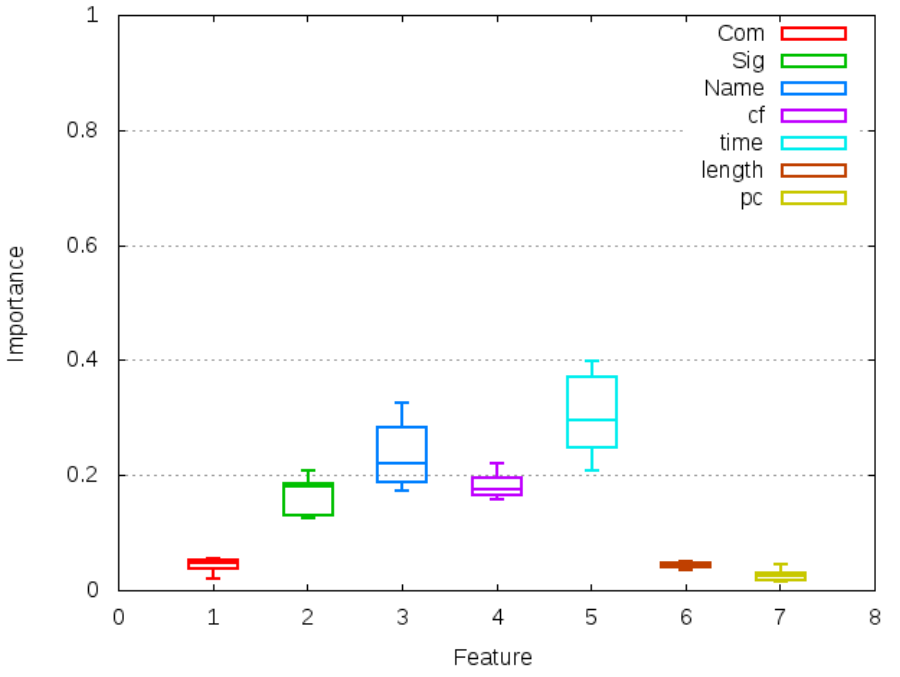


Figure A.33: Feature Importance SWR for cardslib using RF

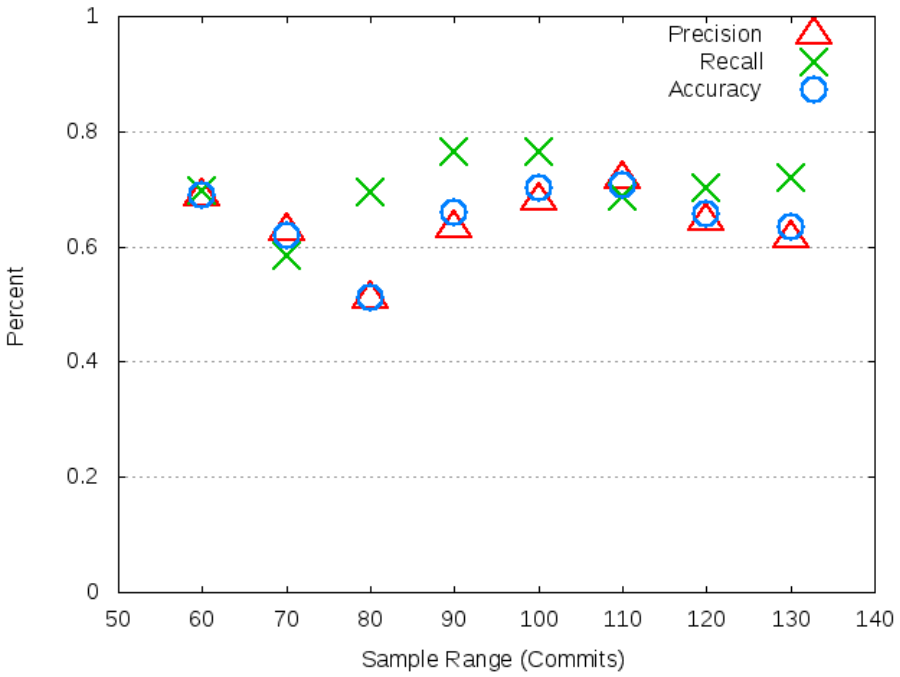


Figure A.34: SWR for dagger using RF

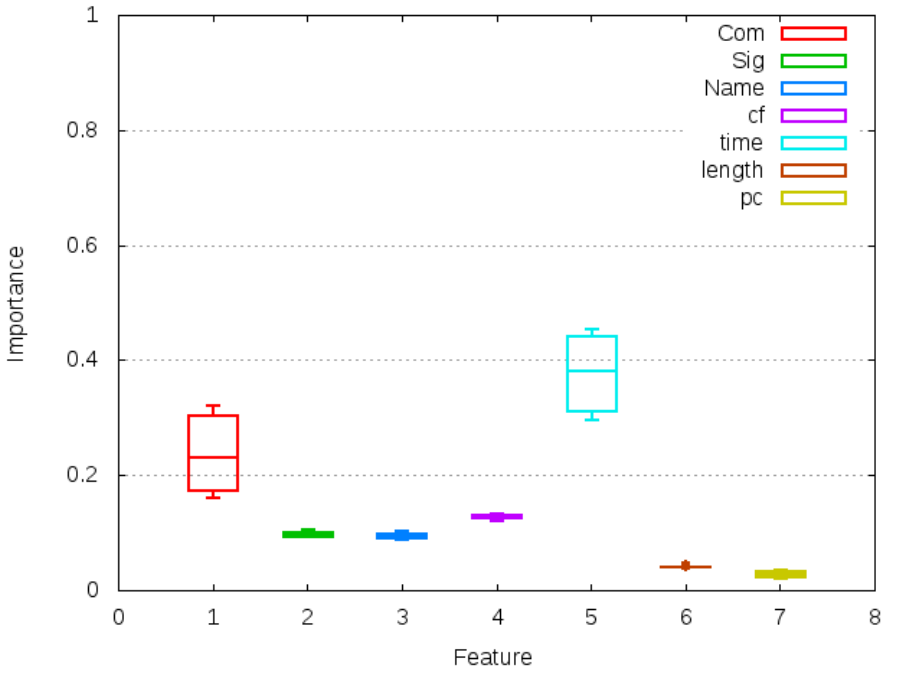


Figure A.35: Feature Importance SWR for dagger using RF

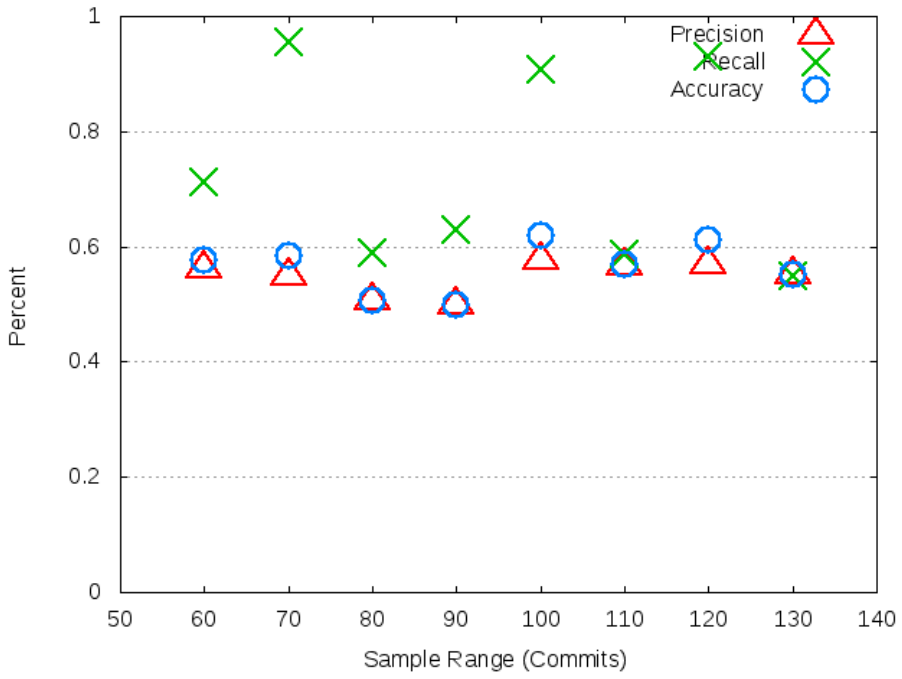


Figure A.36: SWR for deeplearning4j using RF

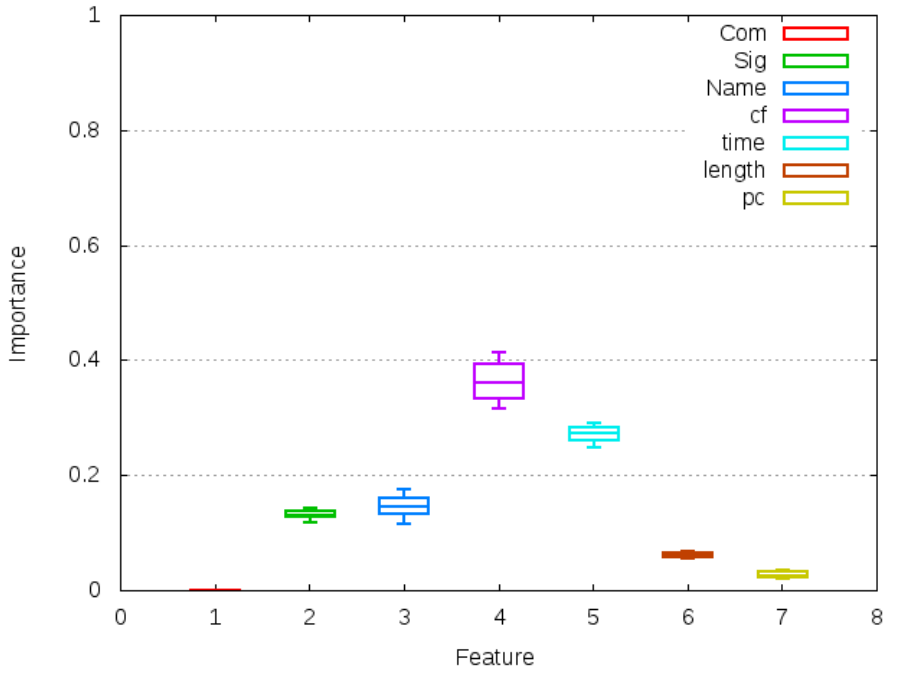


Figure A.37: Feature Importance SWR for deeplearning4j using RF

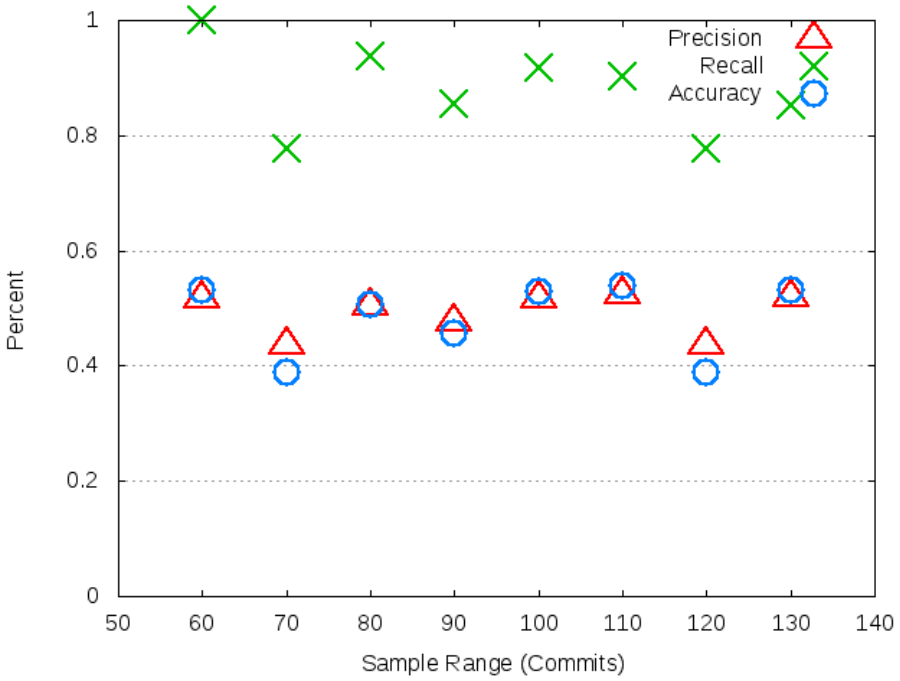


Figure A.38: SWR for fresco using RF

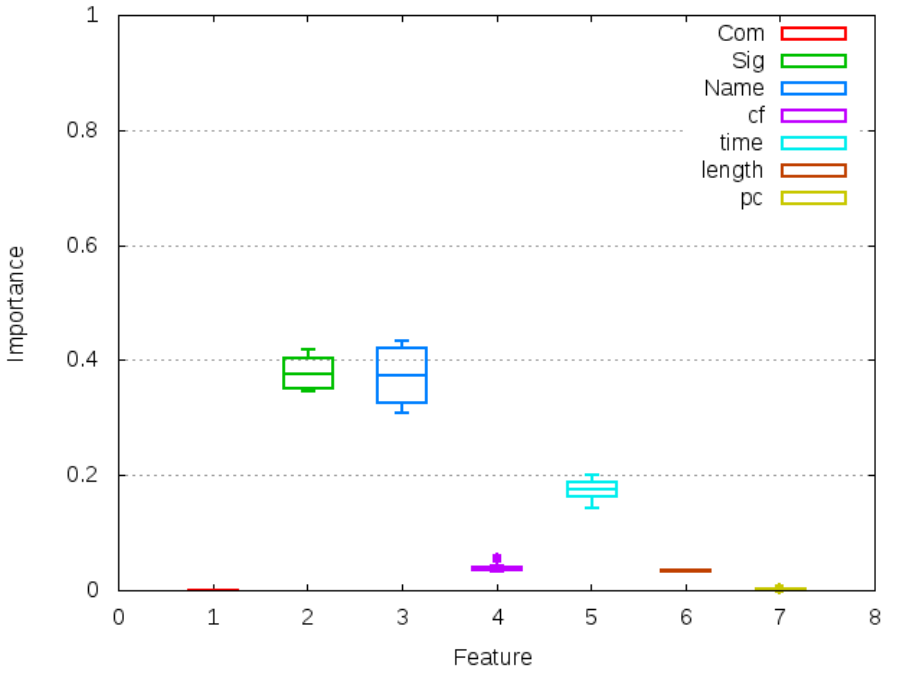


Figure A.39: Feature Importance SWR for fresco using RF

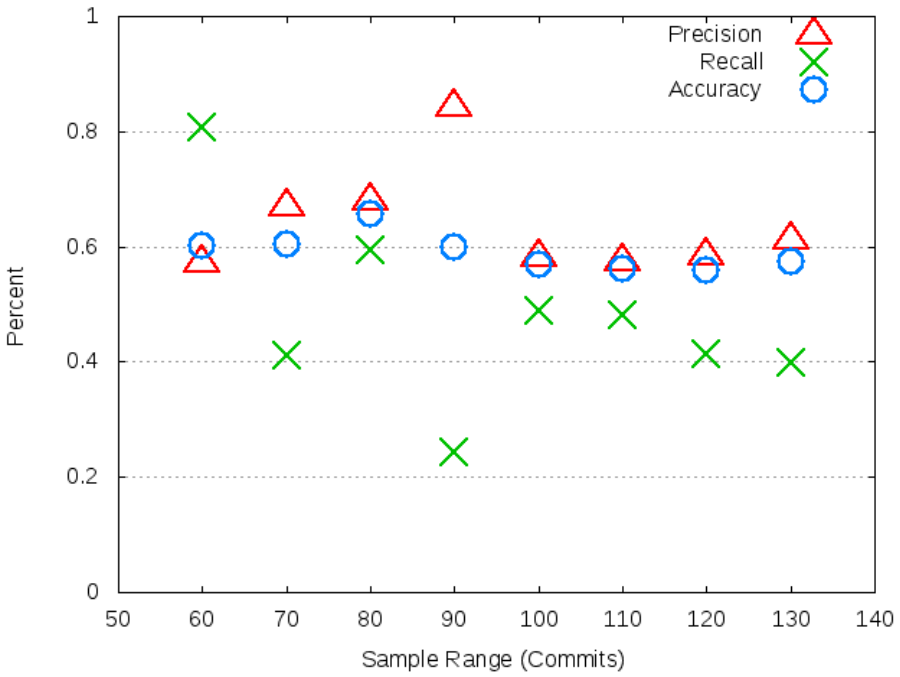


Figure A.40: SWR for governator using RF

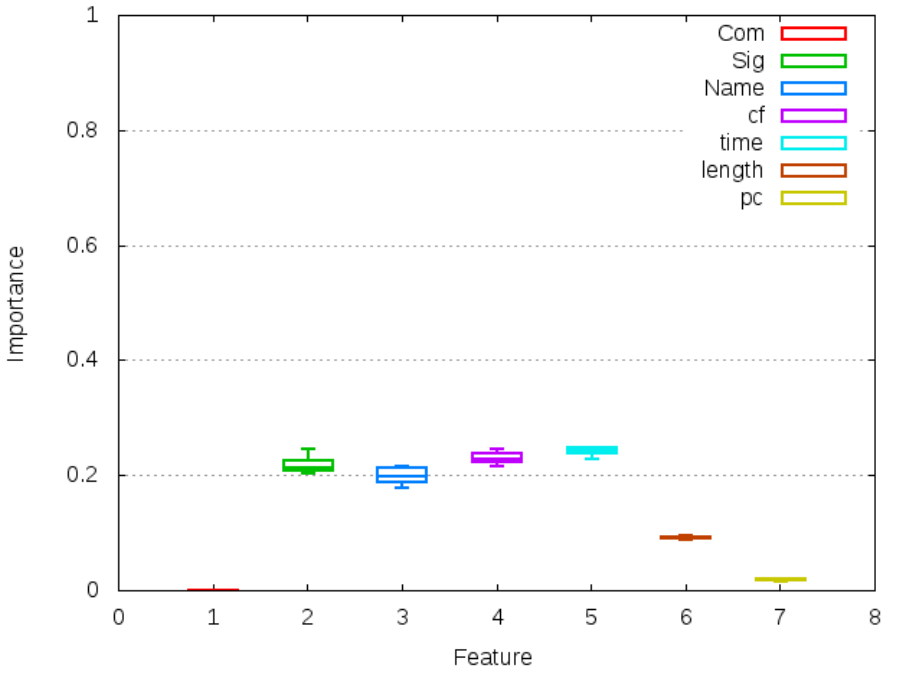


Figure A.41: Feature Importance SWR for governator using RF

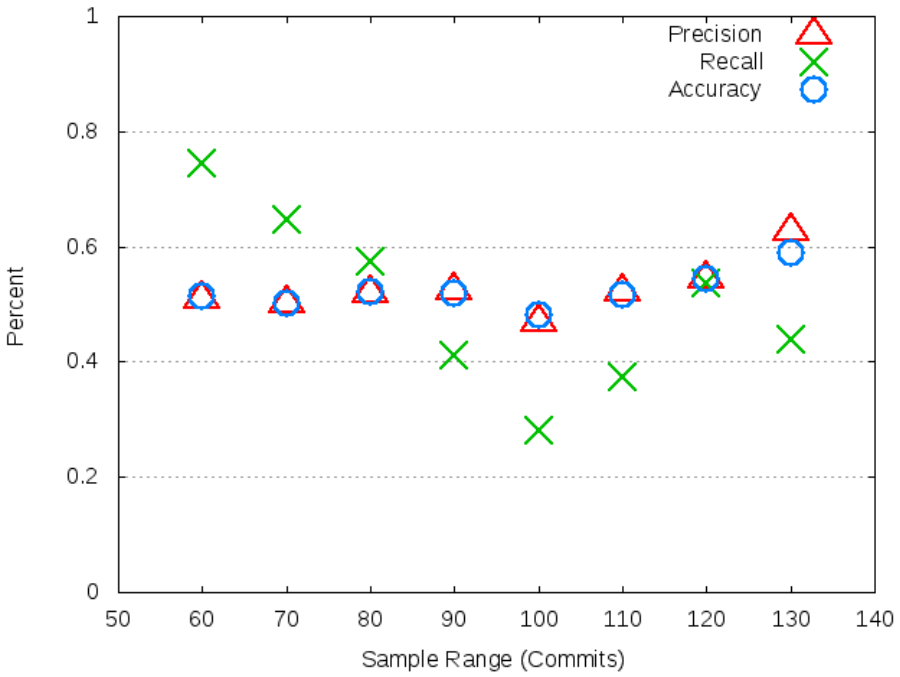


Figure A.42: SWR for greenDAO using RF

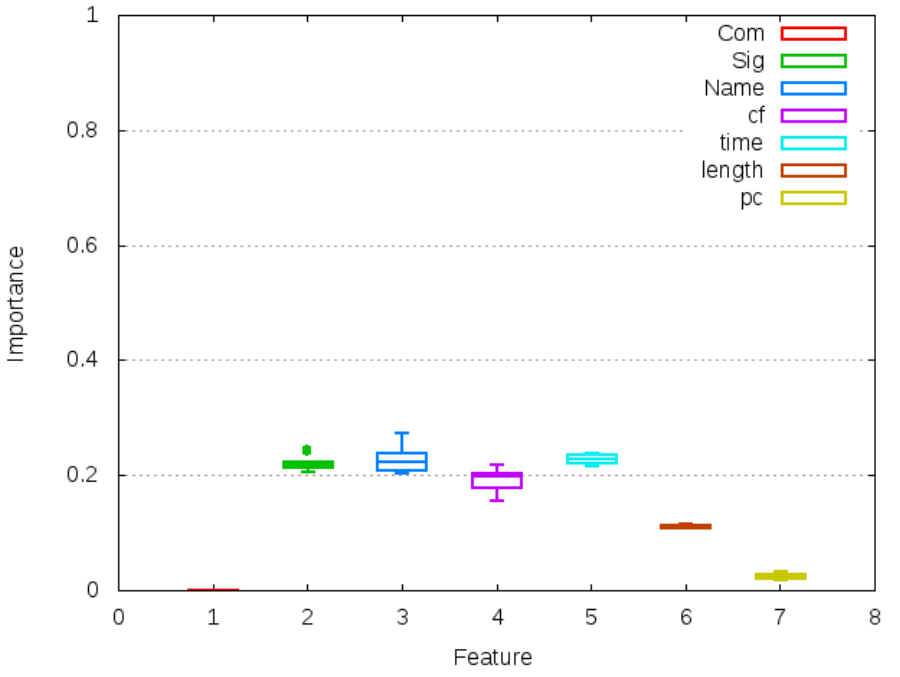


Figure A.43: Feature Importance SWR for greenDAO using RF

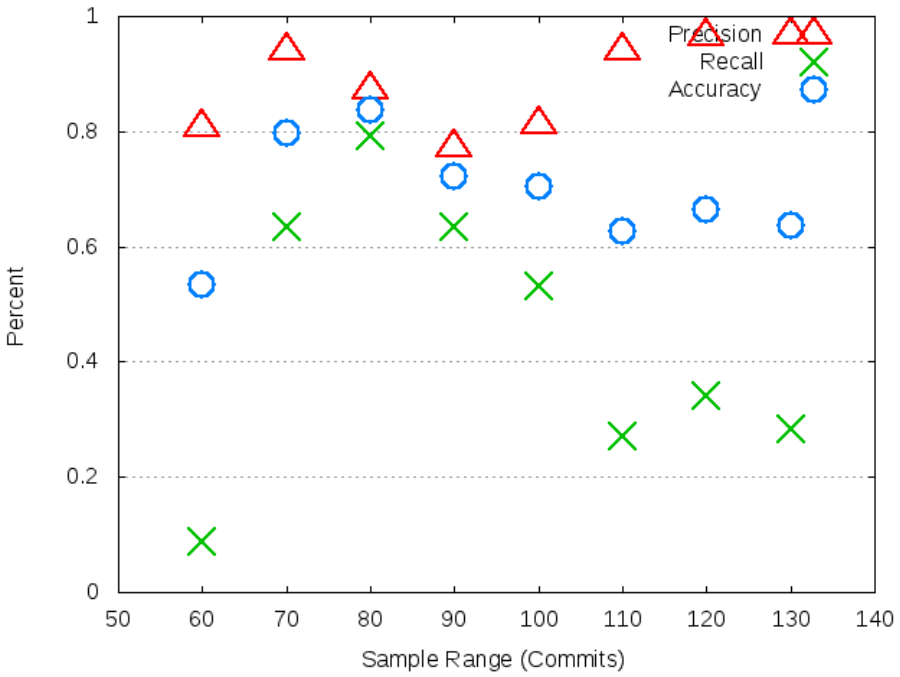


Figure A.44: SWR for http-request using RF

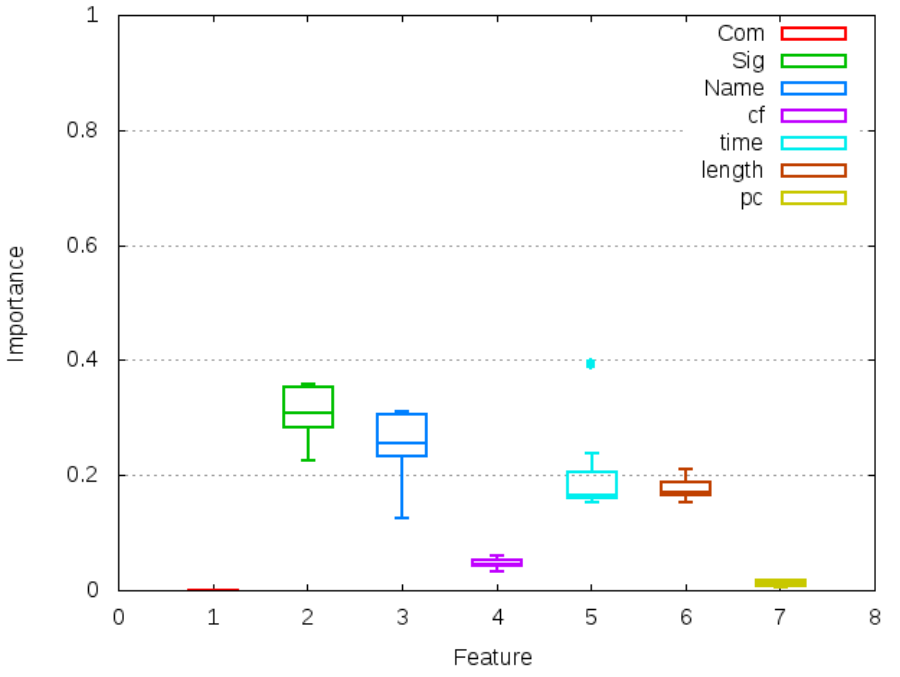


Figure A.45: Feature Importance SWR for http-request using RF

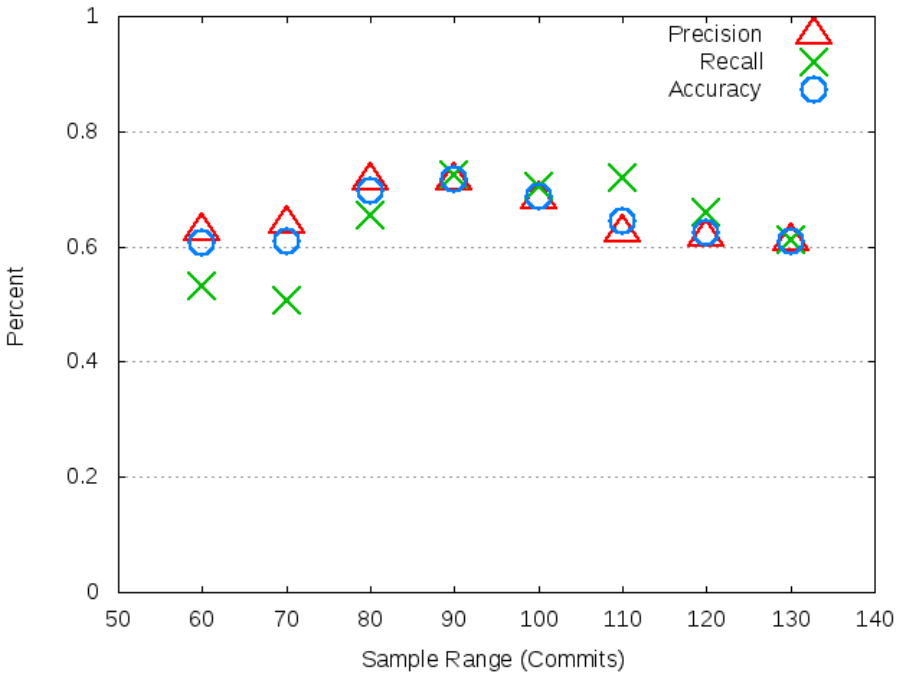


Figure A.46: SWR for ion using RF

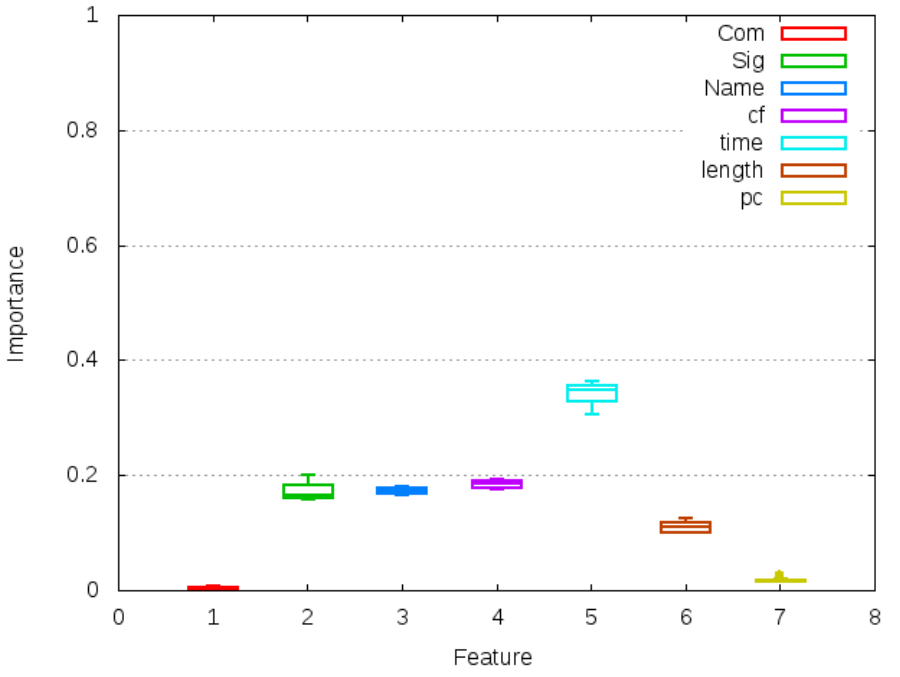


Figure A.47: Feature Importance SWR for ion using RF

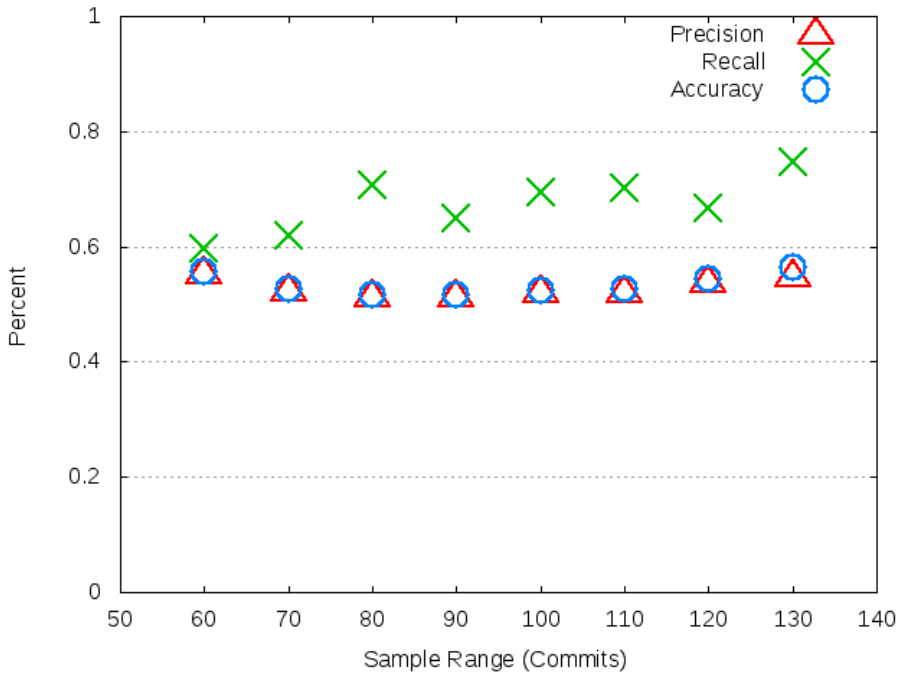


Figure A.48: SWR for jadx using RF

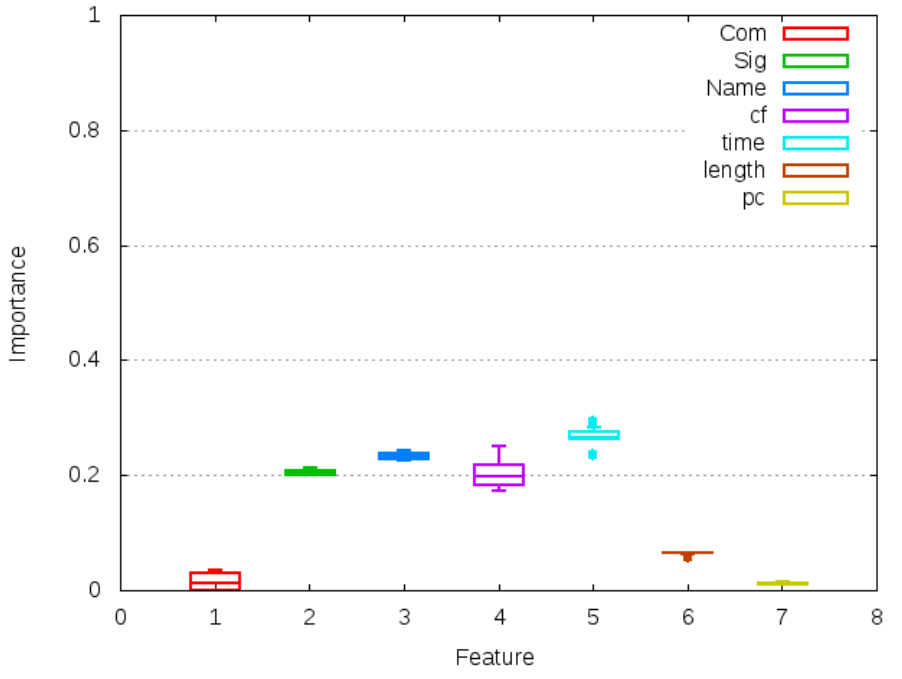


Figure A.49: Feature Importance SWR for jadx using RF

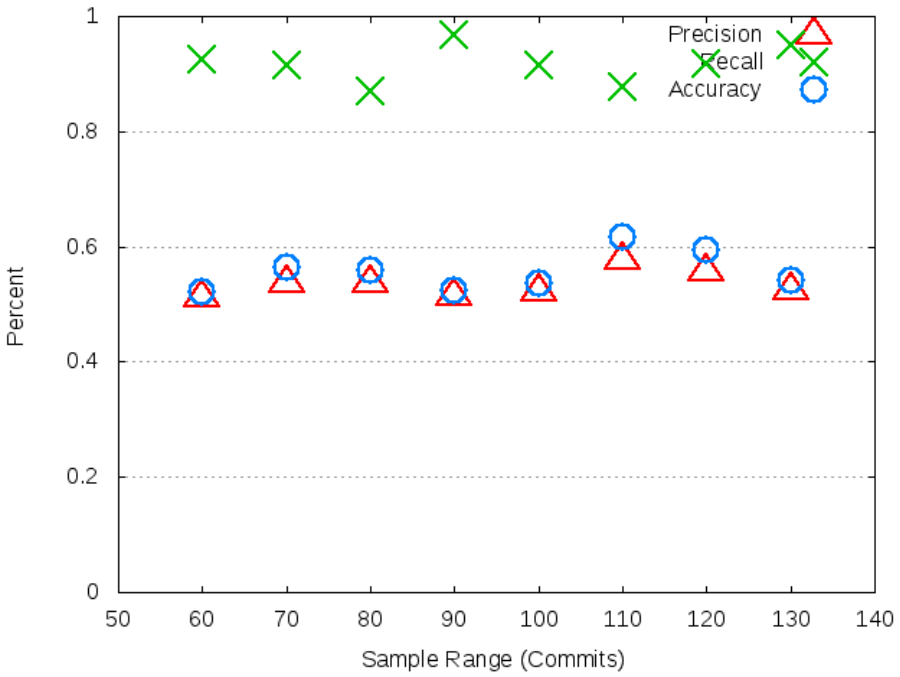


Figure A.50: SWR for mapstruct using RF

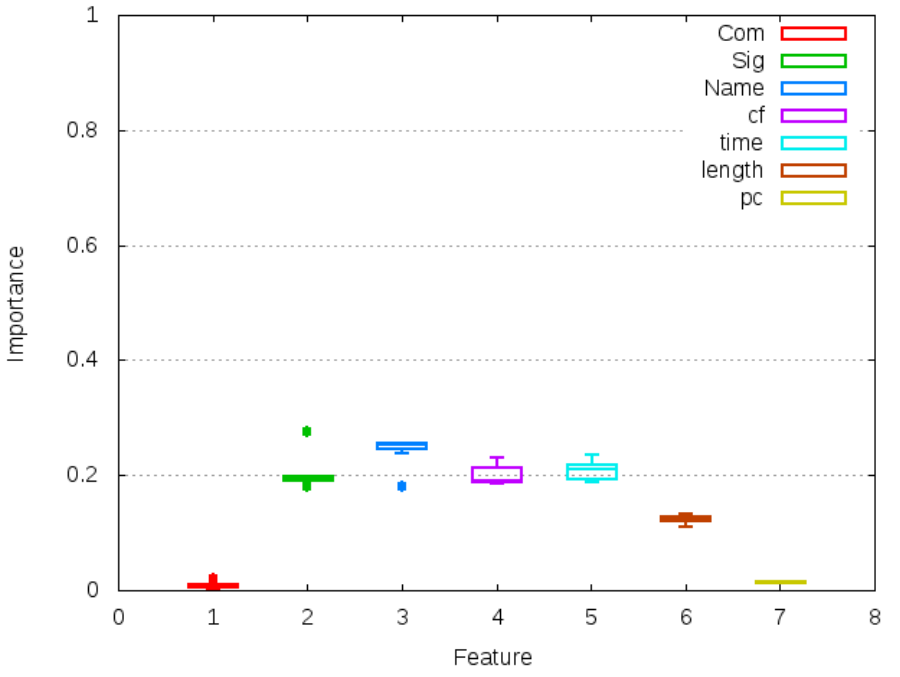


Figure A.51: Feature Importance SWR for mapstruct using RF

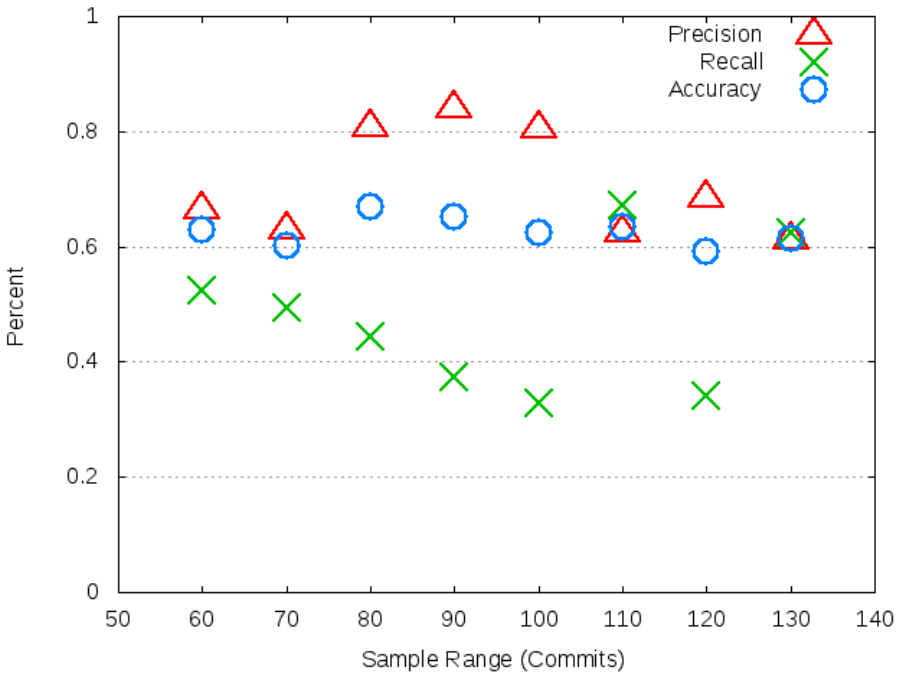


Figure A.52: SWR for nettosphere using RF

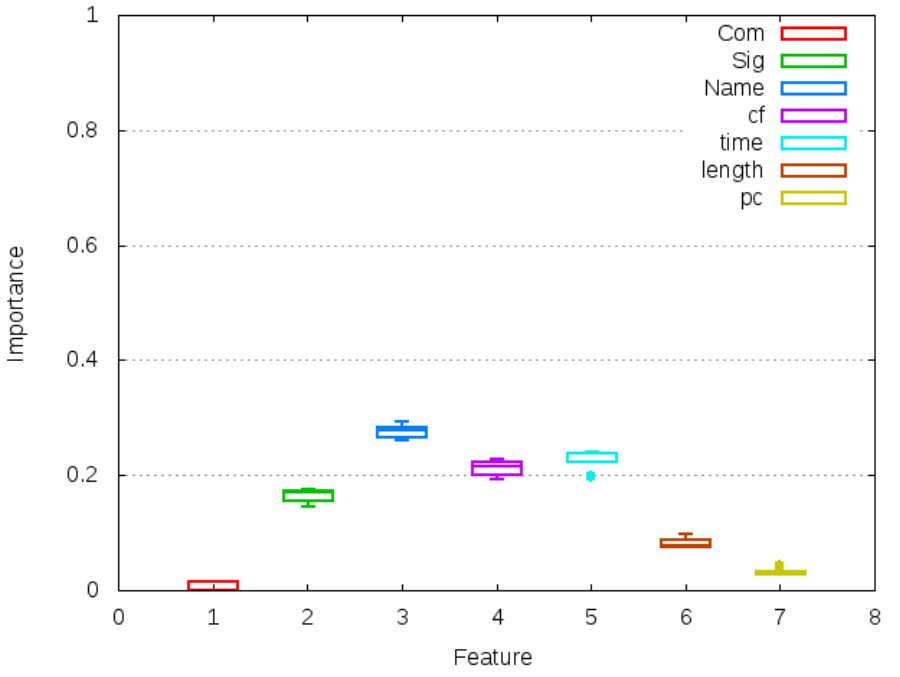


Figure A.53: Feature Importance SWR for nettosphere using RF

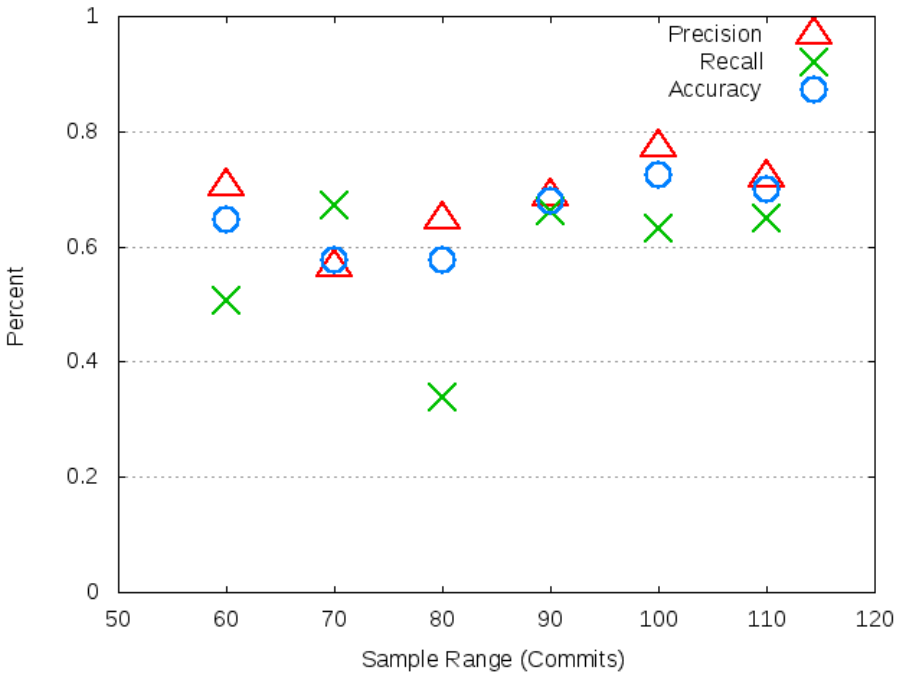


Figure A.54: SWR for parceler using RF

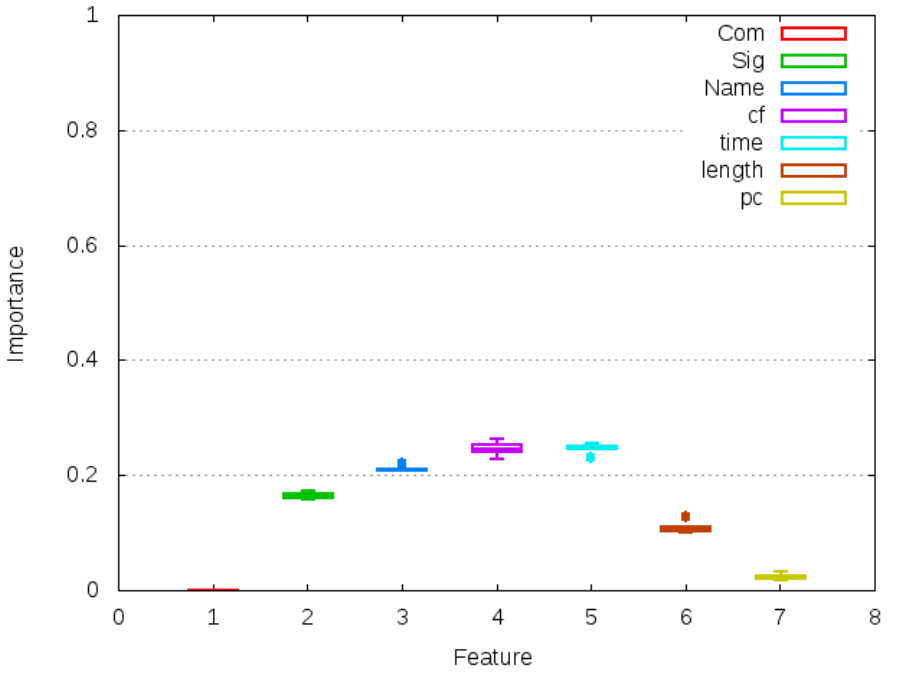


Figure A.55: Feature Importance SWR for parceler using RF

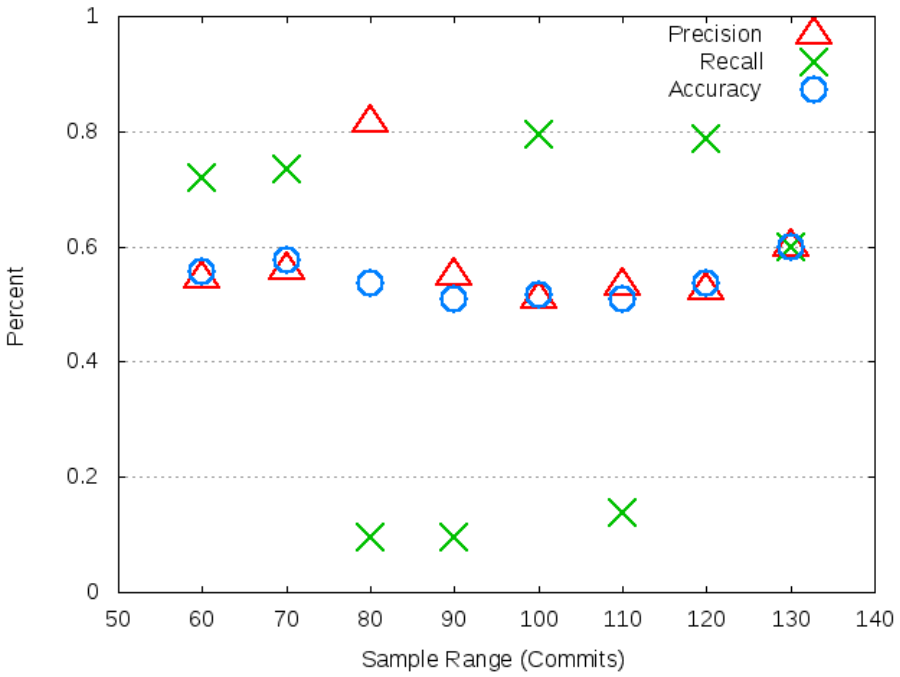


Figure A.56: SWR for retrolambda using RF

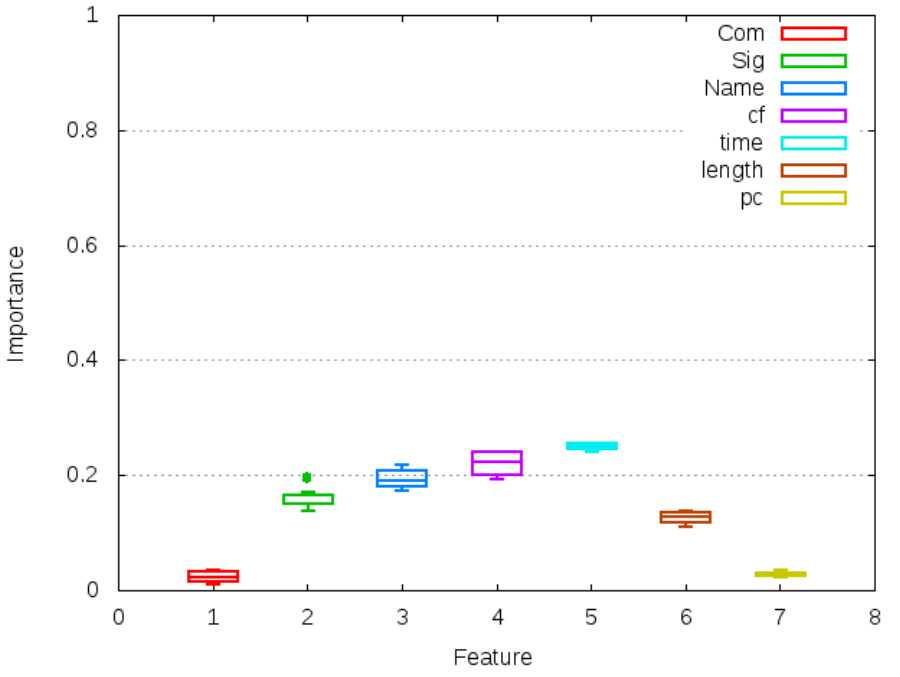


Figure A.57: Feature Importance SWR for retrolambda using RF

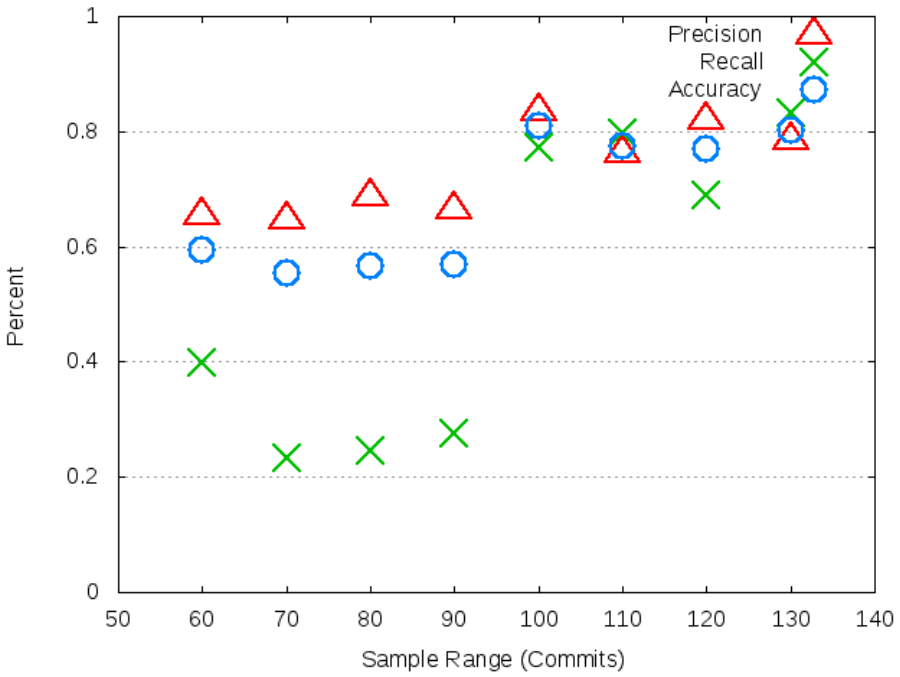


Figure A.58: SWR for ShowcaseView using RF

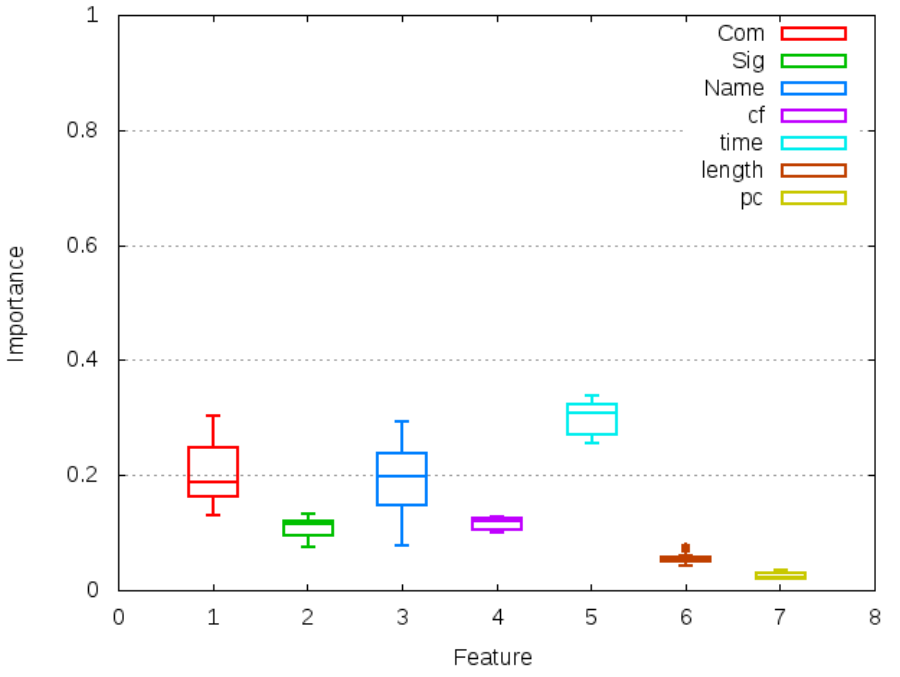


Figure A.59: Feature Importance SWR for ShowcaseView using RF

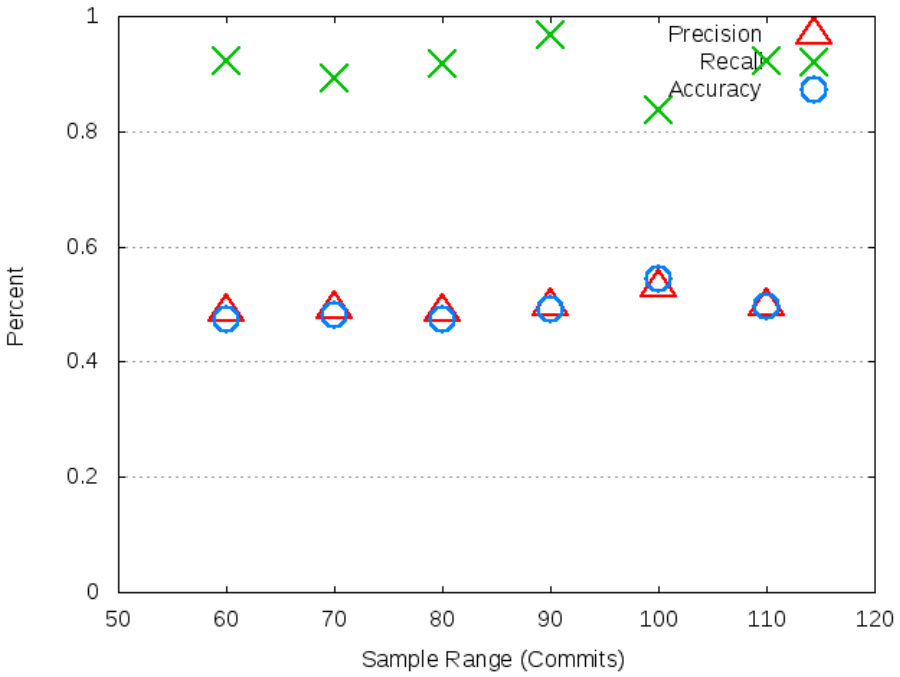


Figure A.60: SWR for smile using RF

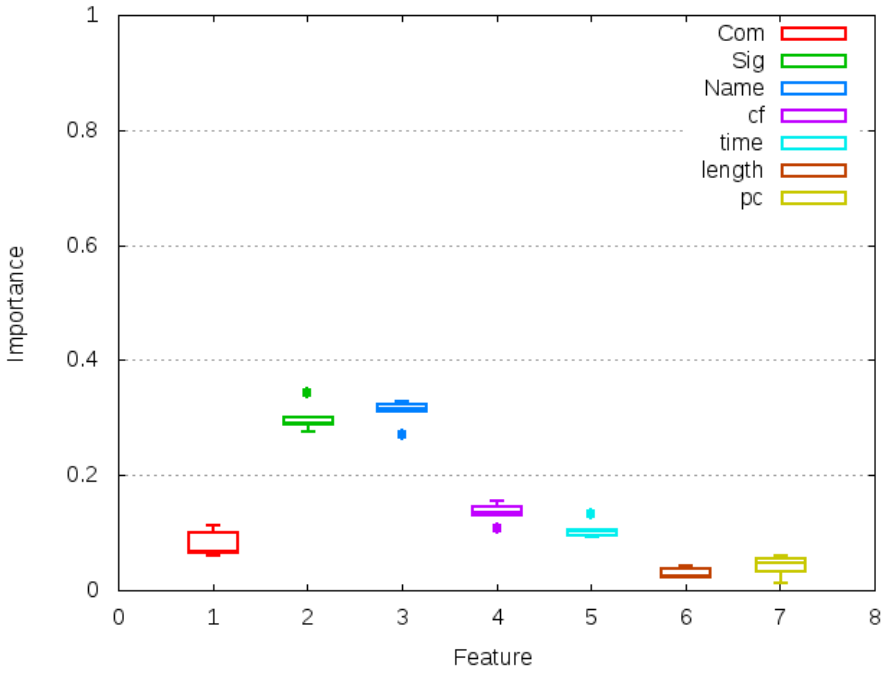


Figure A.61: Feature Importance SWR for smile using RF

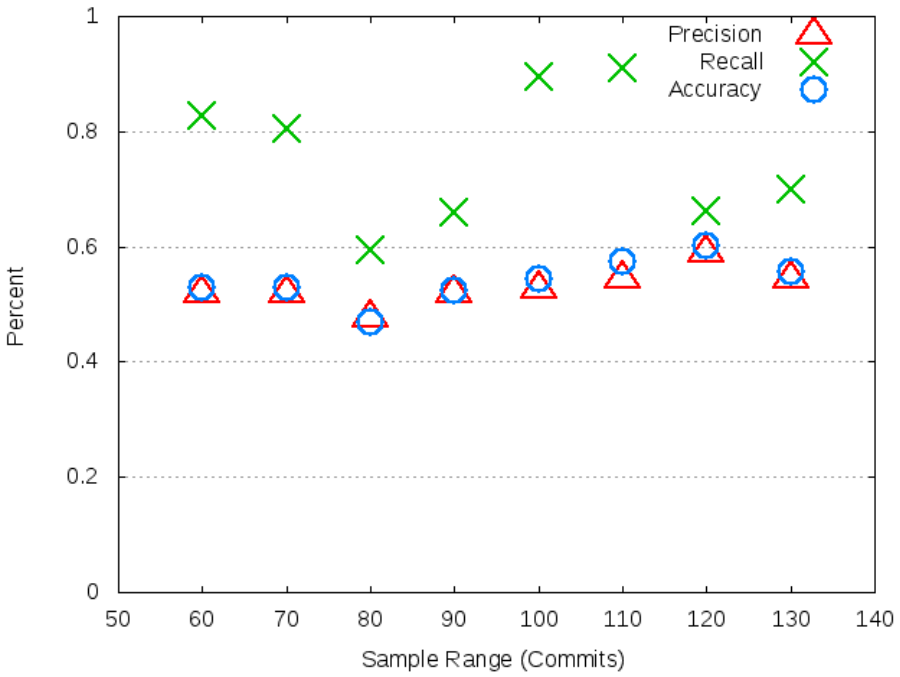


Figure A.62: SWR for spark using RF

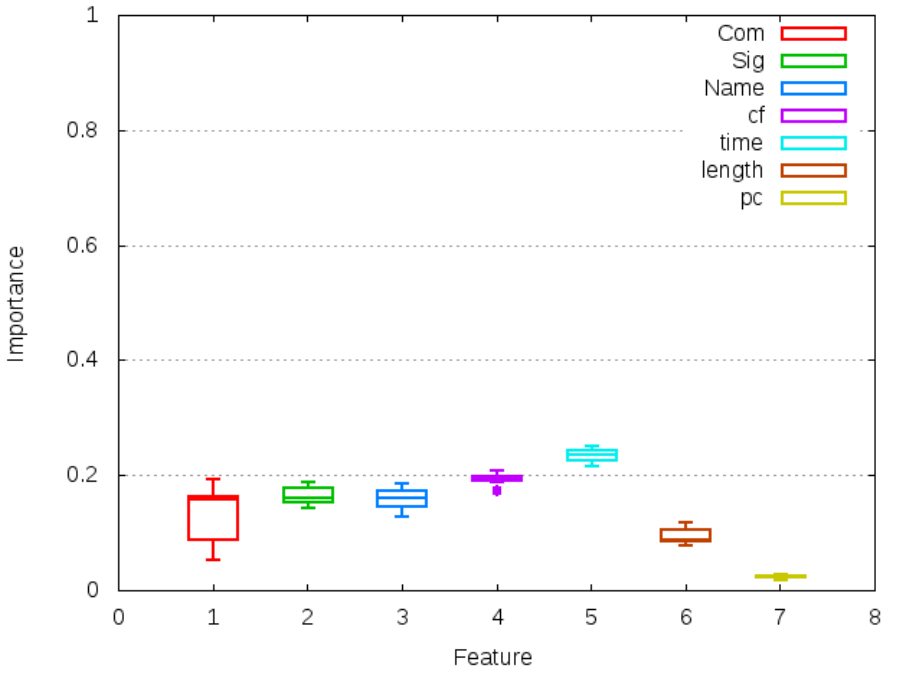


Figure A.63: Feature Importance SWR for spark using RF

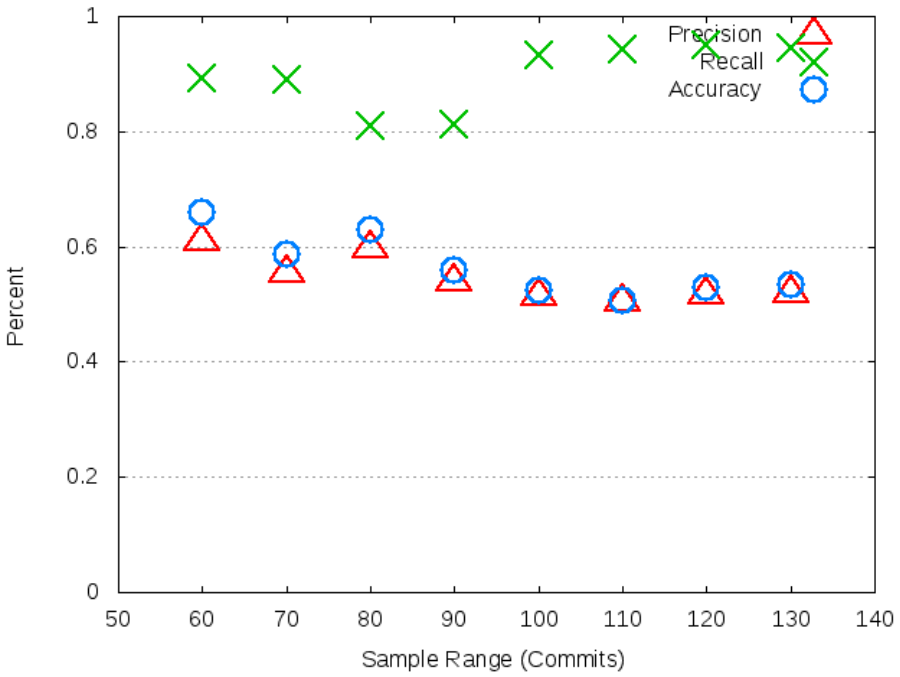


Figure A.64: SWR for storm using RF

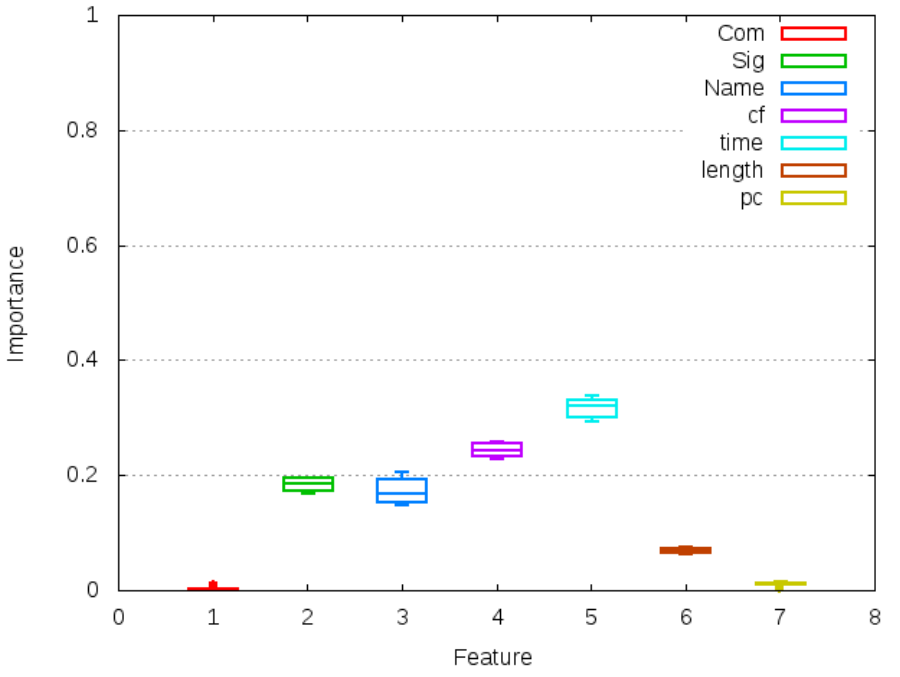


Figure A.65: Feature Importance SWR for storm using RF

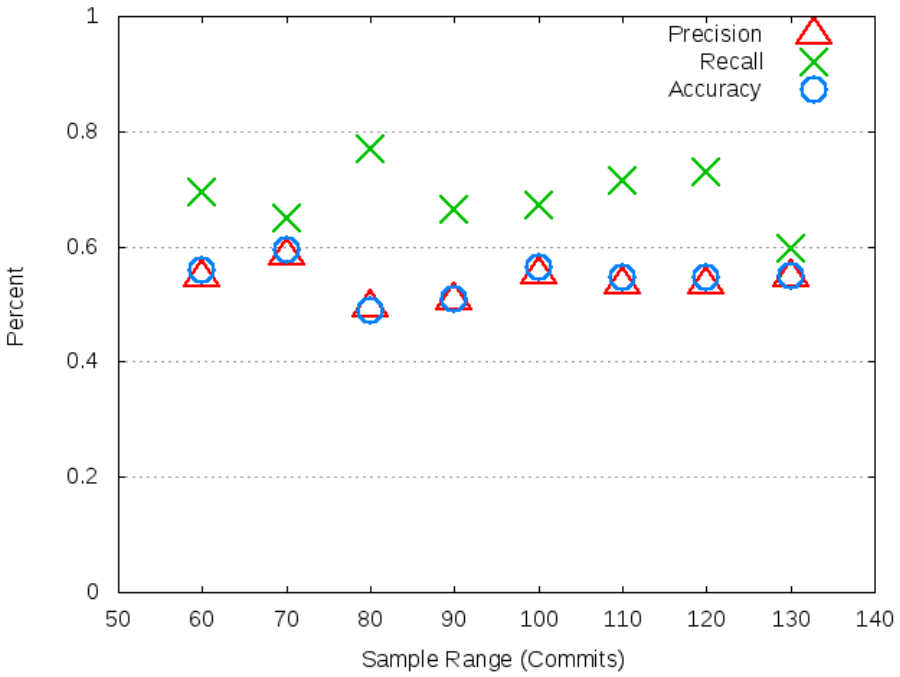


Figure A.66: SWR for tempto using RF

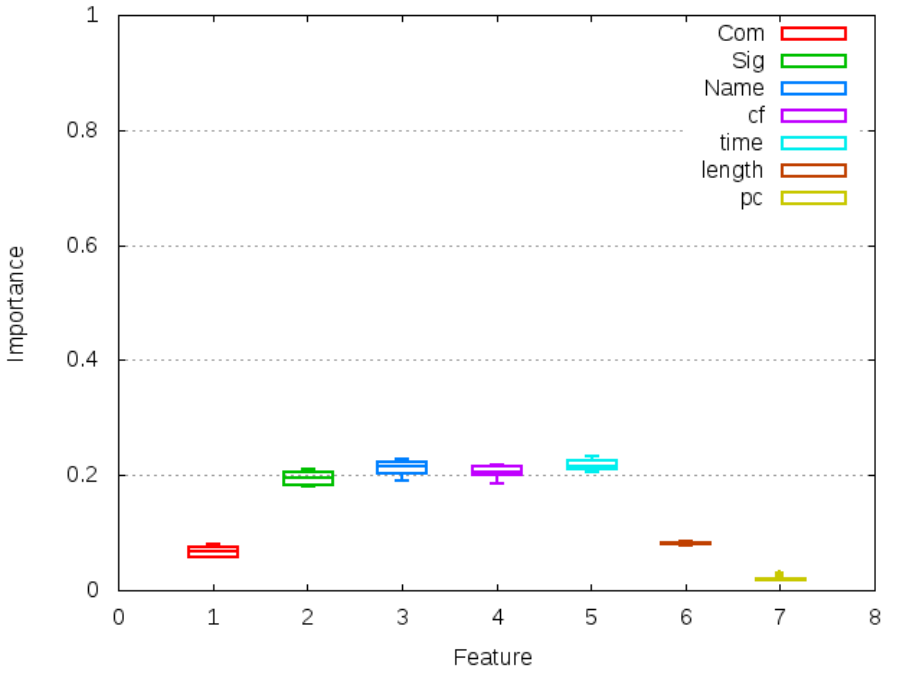


Figure A.67: Feature Importance SWR for tempto using RF

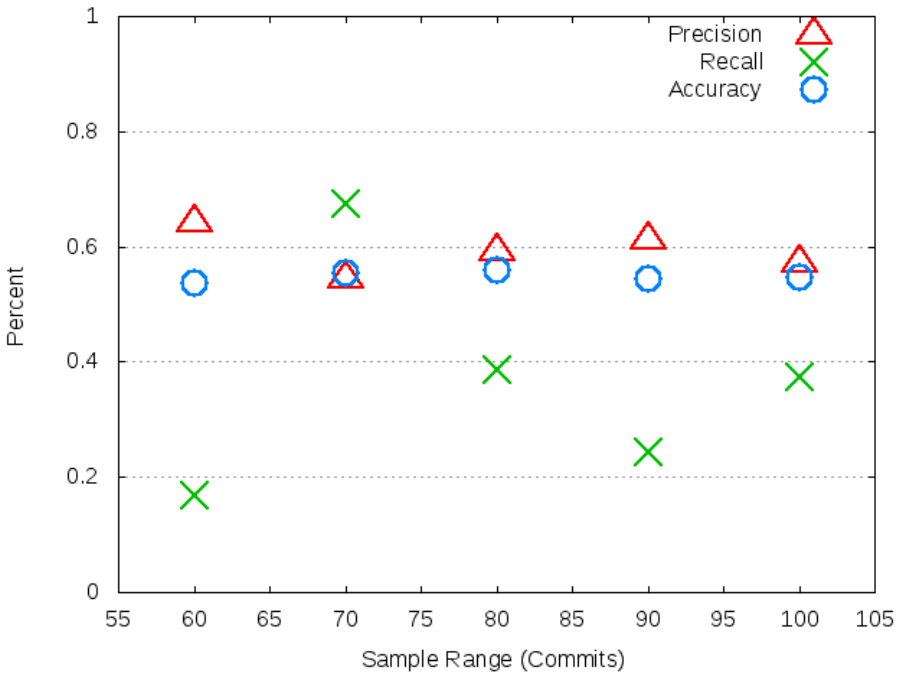


Figure A.68: SWR for yardstick using RF

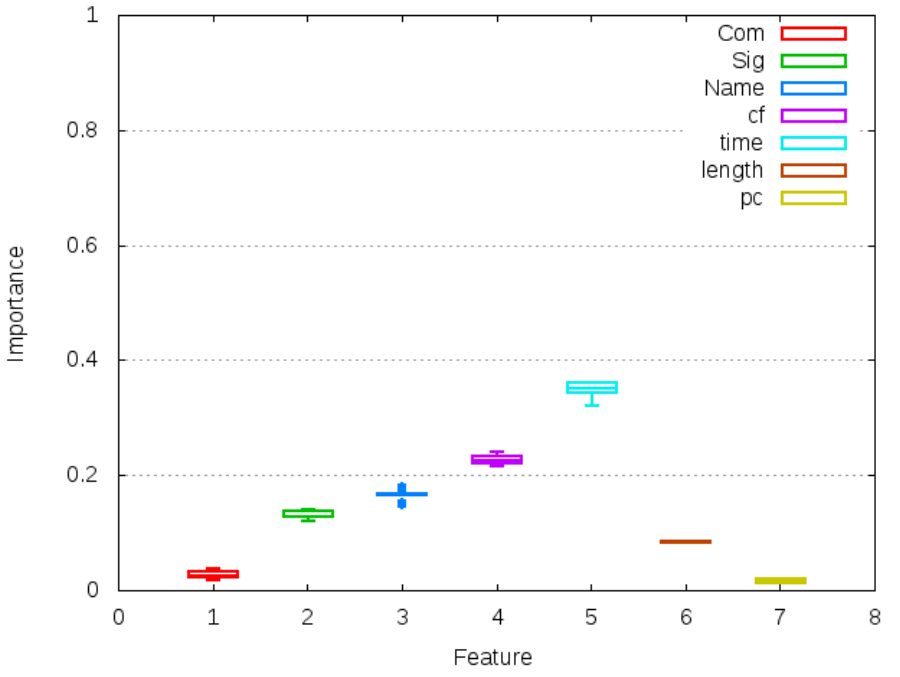


Figure A.69: Feature Importance SWR for yardstick using RF

A.2 Experiment 2

A.2.1 Support Vector Machine

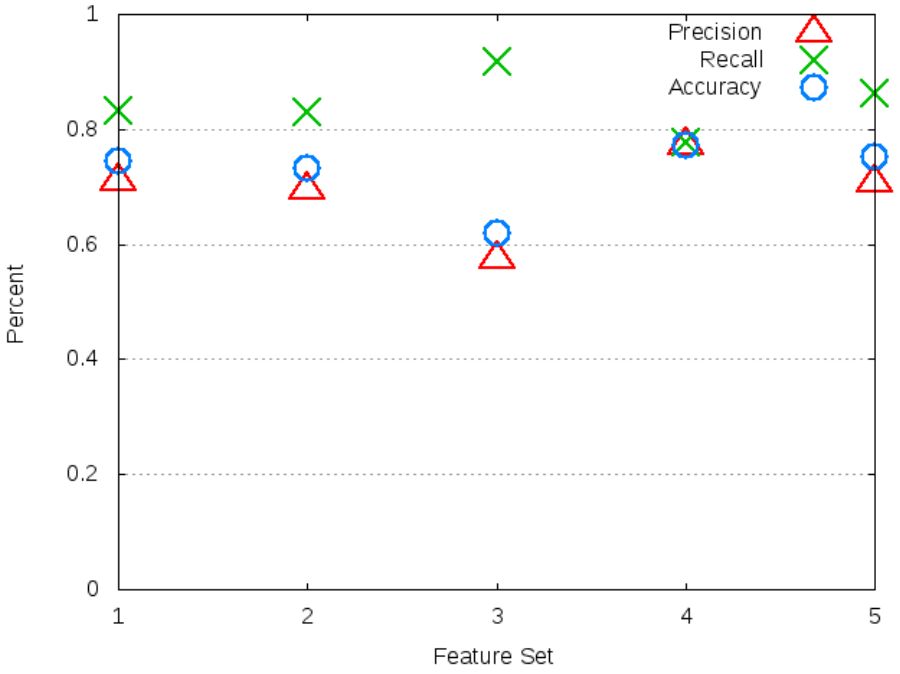


Figure A.70: Feature for acra using SVM

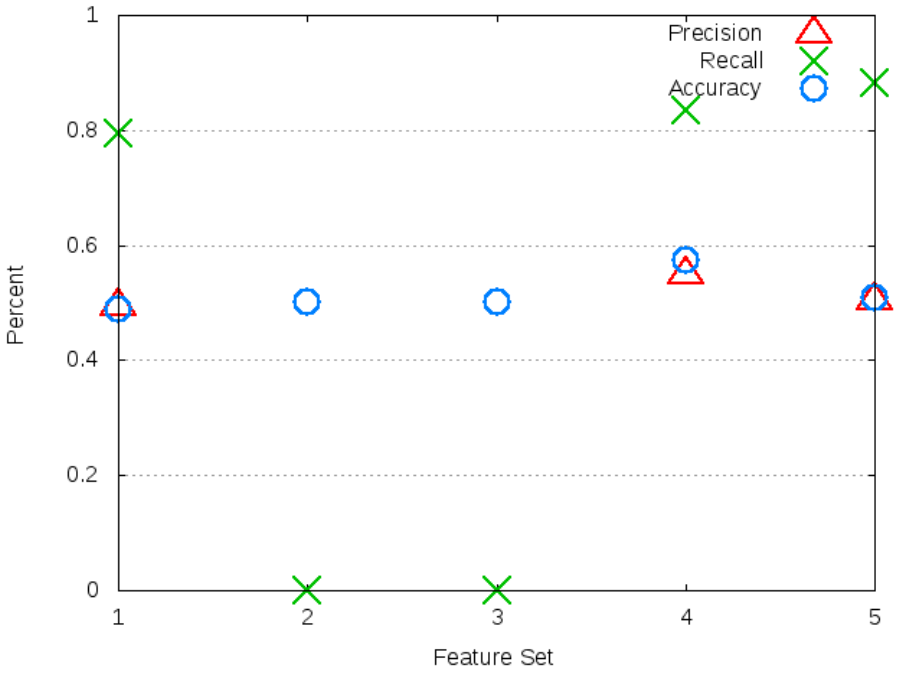


Figure A.71: Feature for arquillian-core using SVM

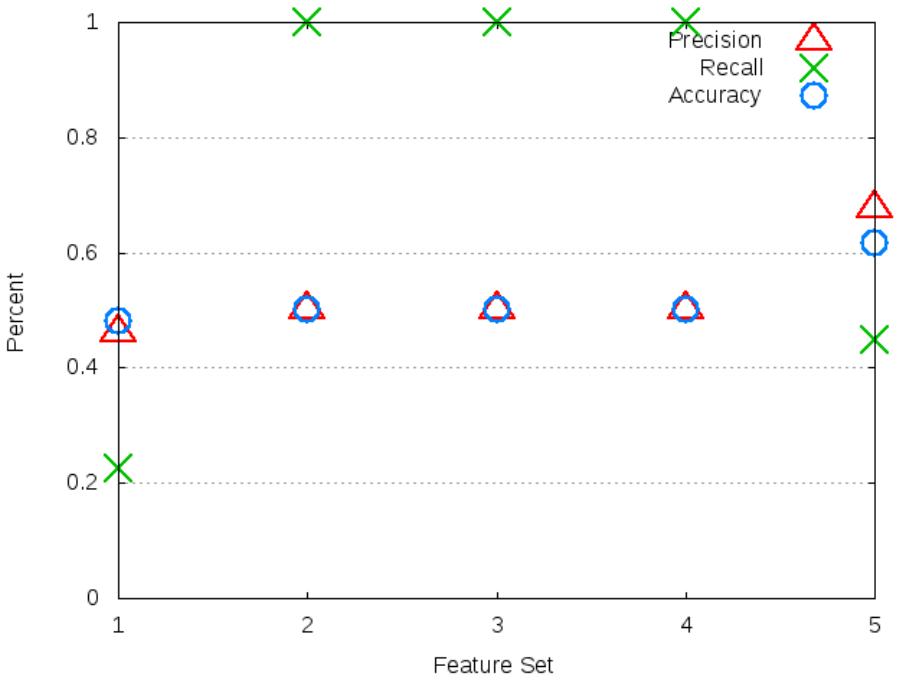


Figure A.72: Feature for blockly-android using SVM

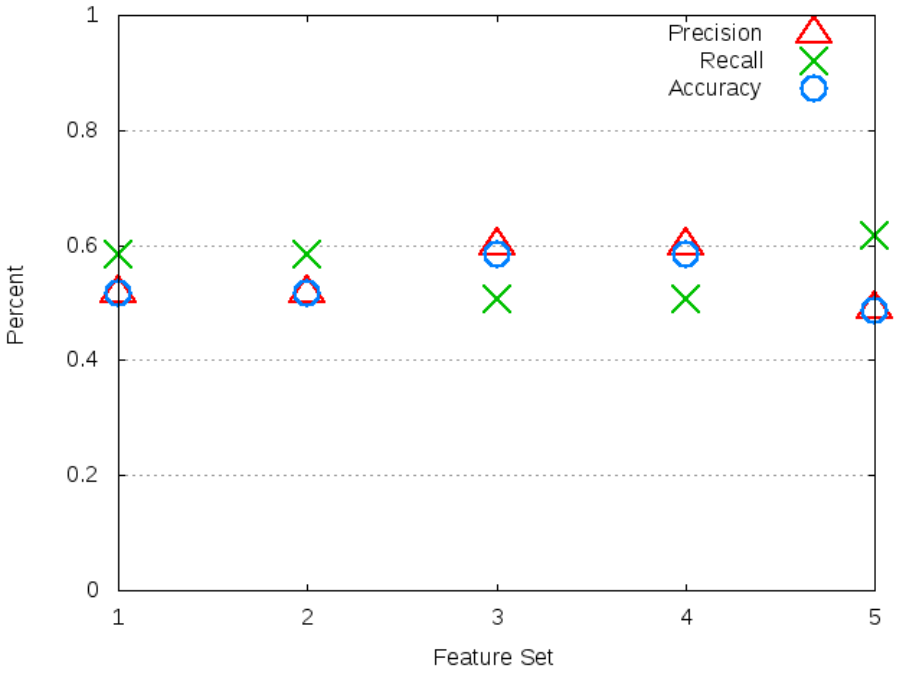


Figure A.73: Feature for brave using SVM

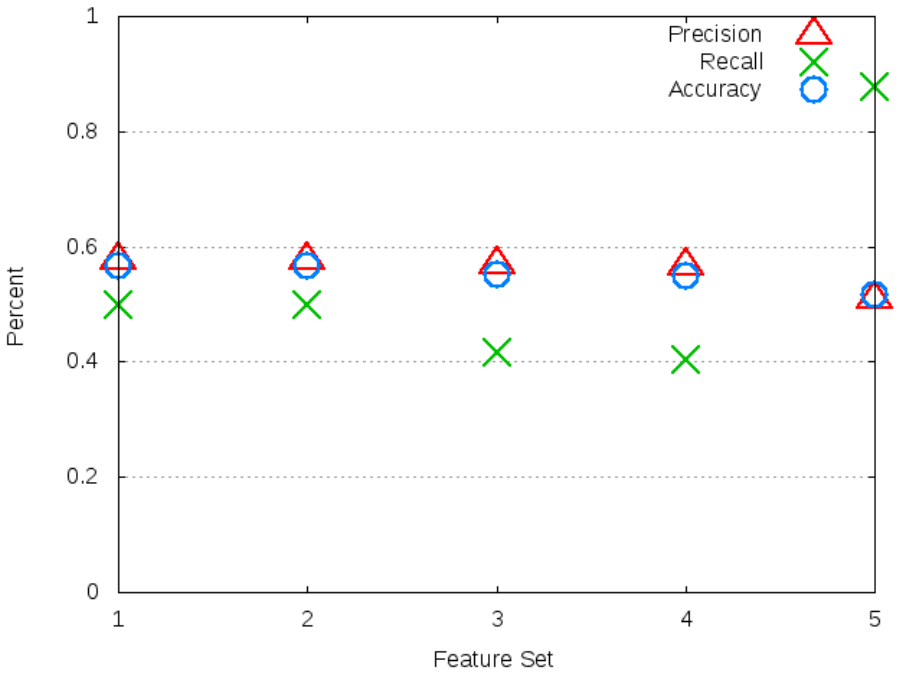


Figure A.74: Feature for cardslib using SVM

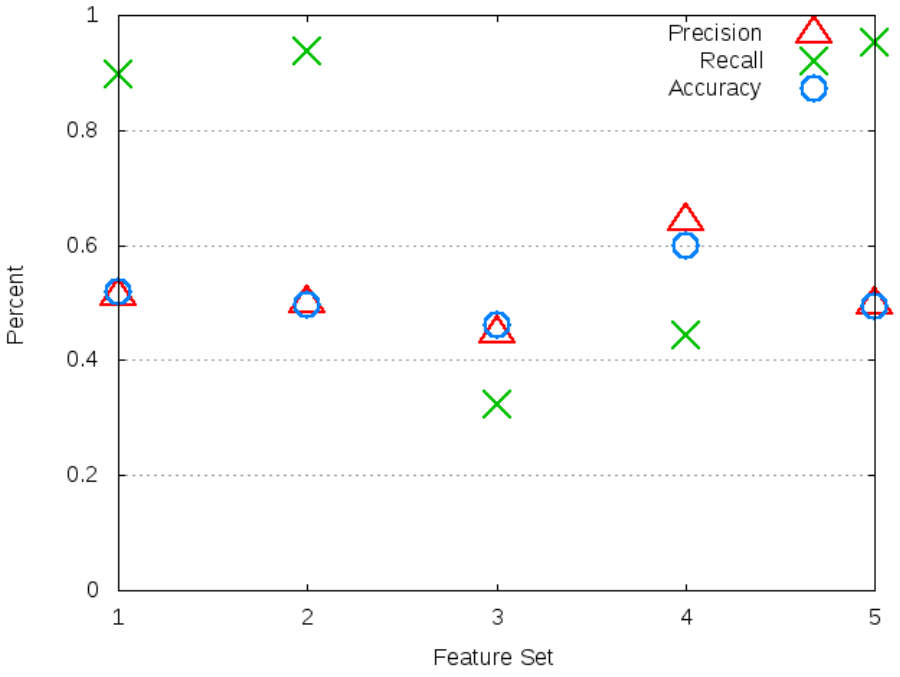


Figure A.75: Feature for dagger using SVM

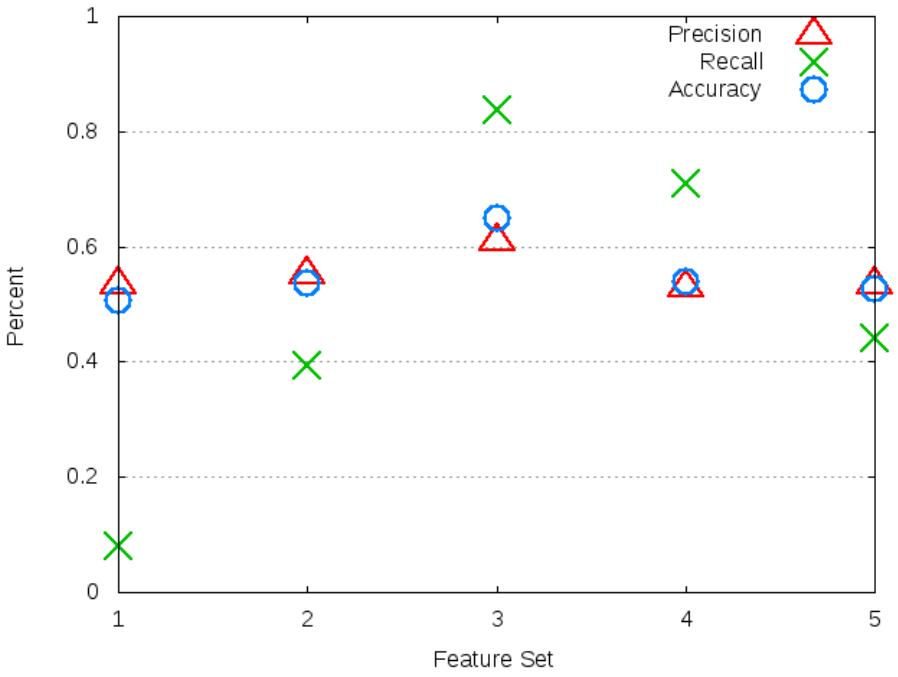


Figure A.76: Feature for deeplearning4j using SVM

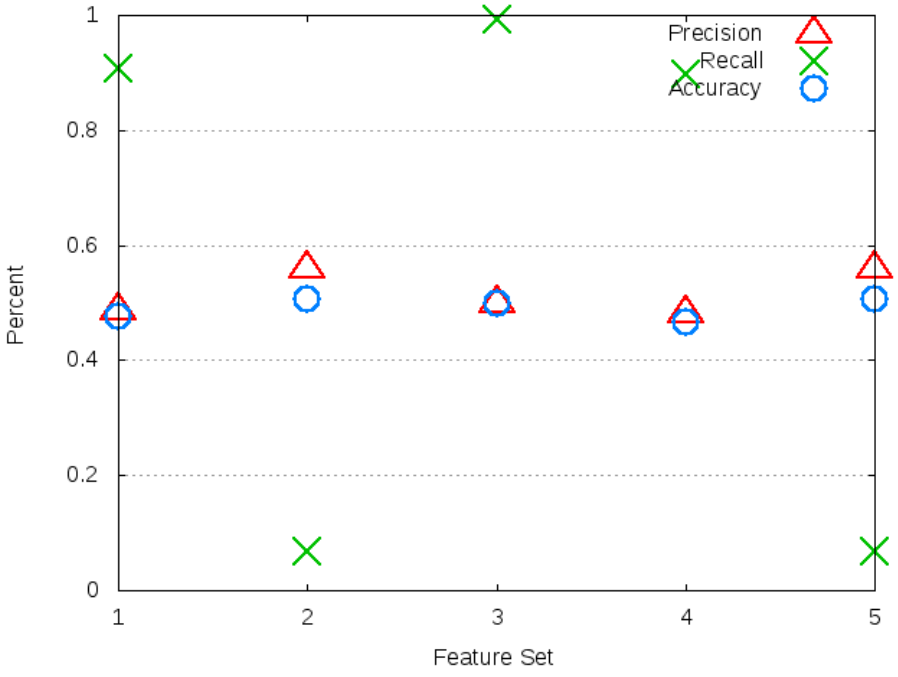


Figure A.77: Feature for fresco using SVM

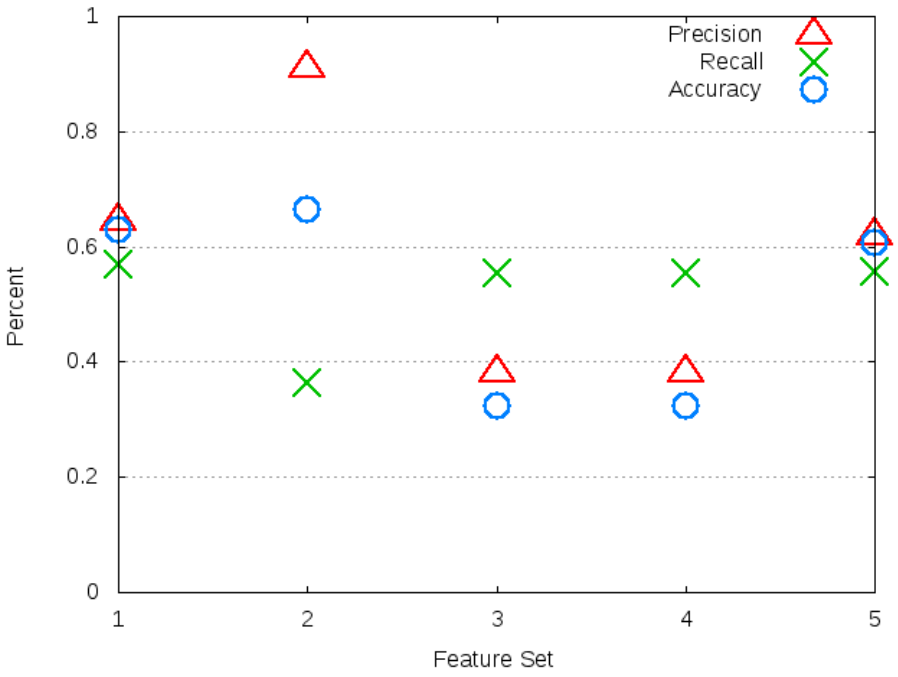


Figure A.78: Feature for governator using SVM

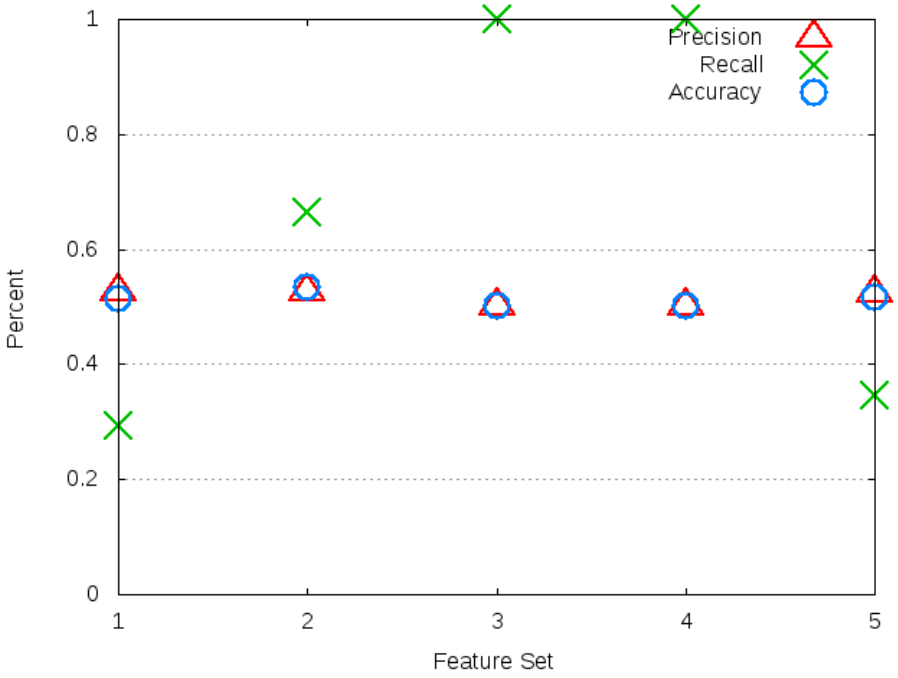


Figure A.79: Feature for greenDAO using SVM

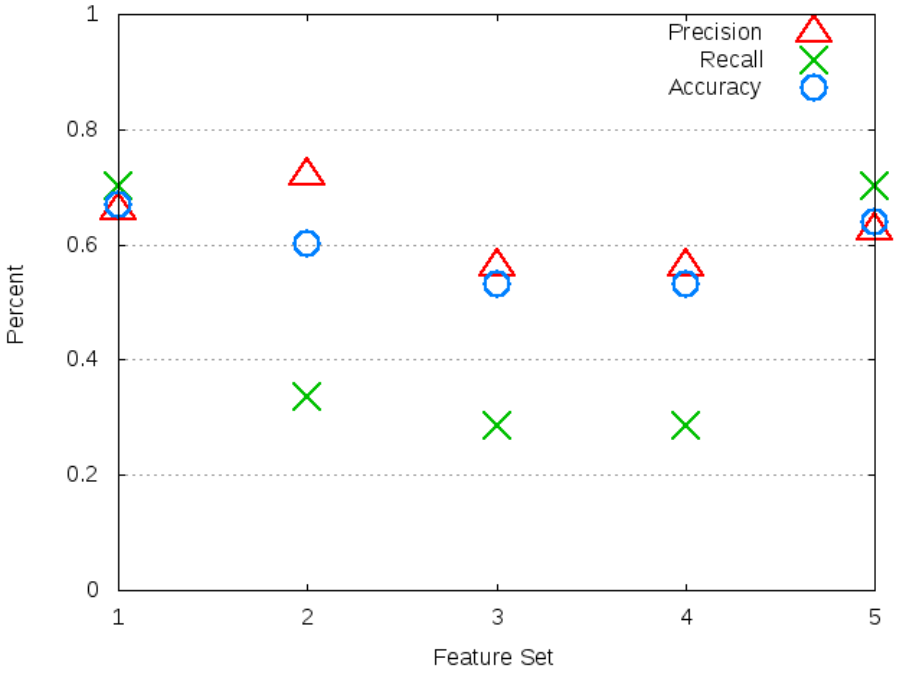


Figure A.80: Feature for http-request using SVM

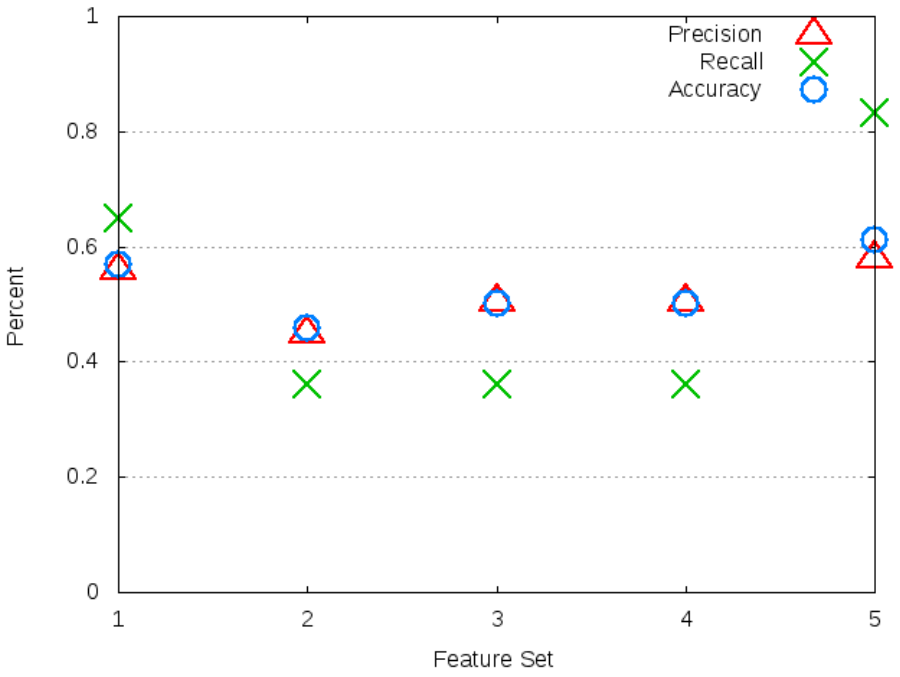


Figure A.81: Feature for ion using SVM

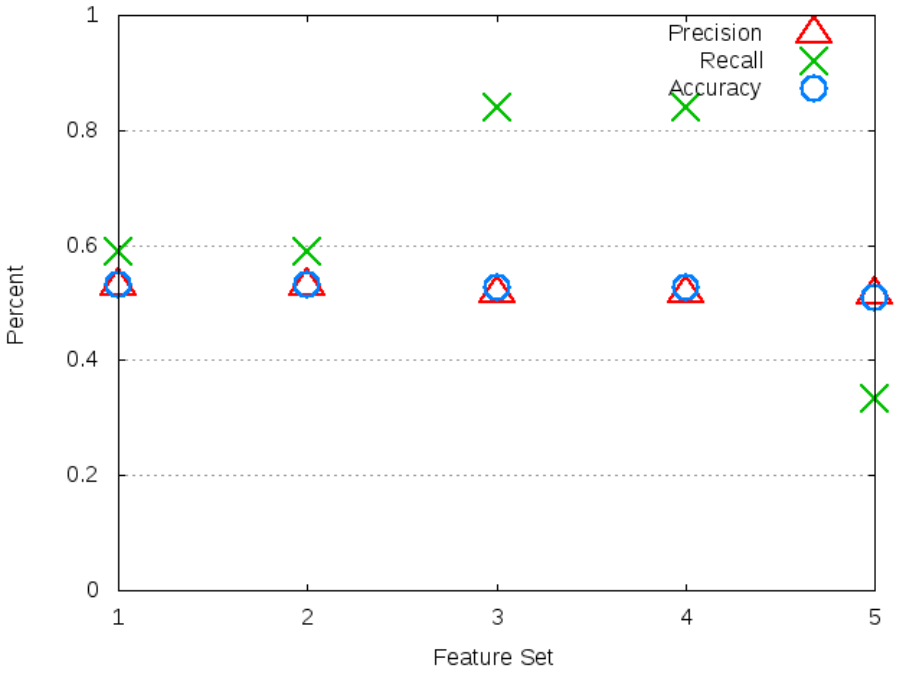


Figure A.82: Feature for jadx using SVM

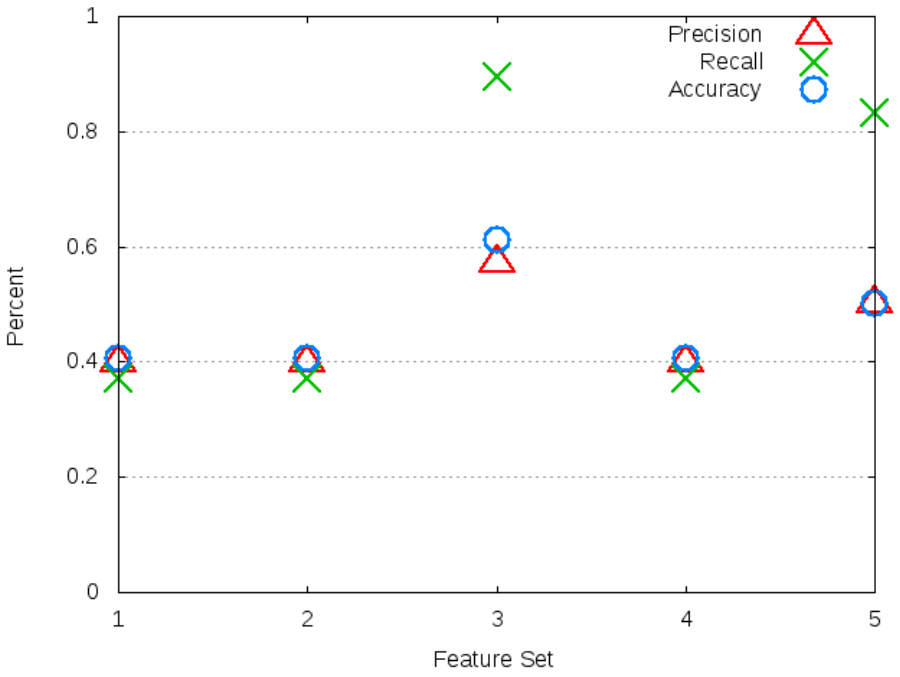


Figure A.83: Feature for mapstruct using SVM

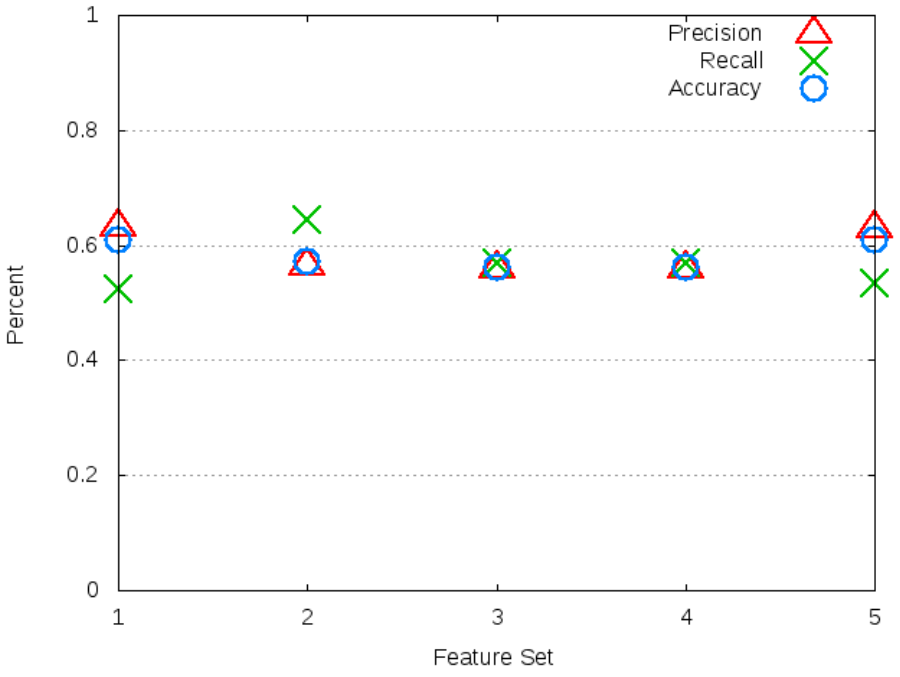


Figure A.84: Feature for nettosphere using SVM

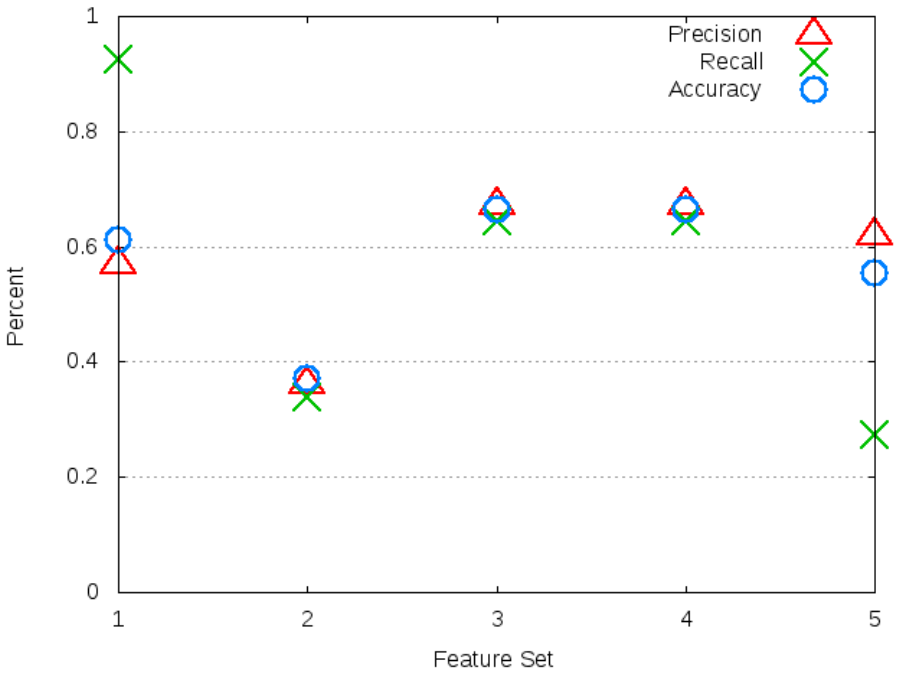


Figure A.85: Feature for parceler using SVM

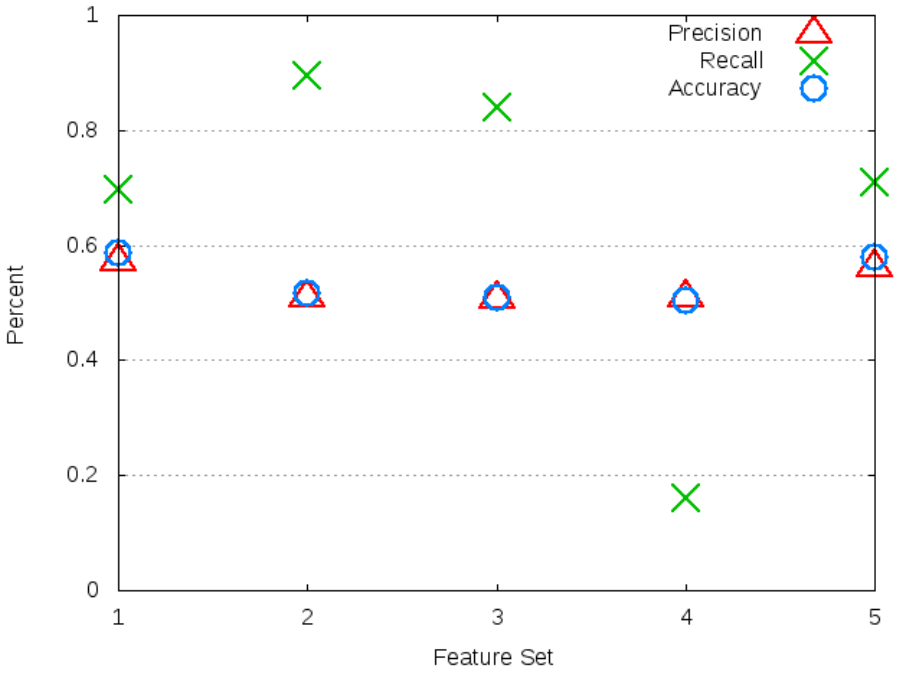


Figure A.86: Feature for retrolambda using SVM

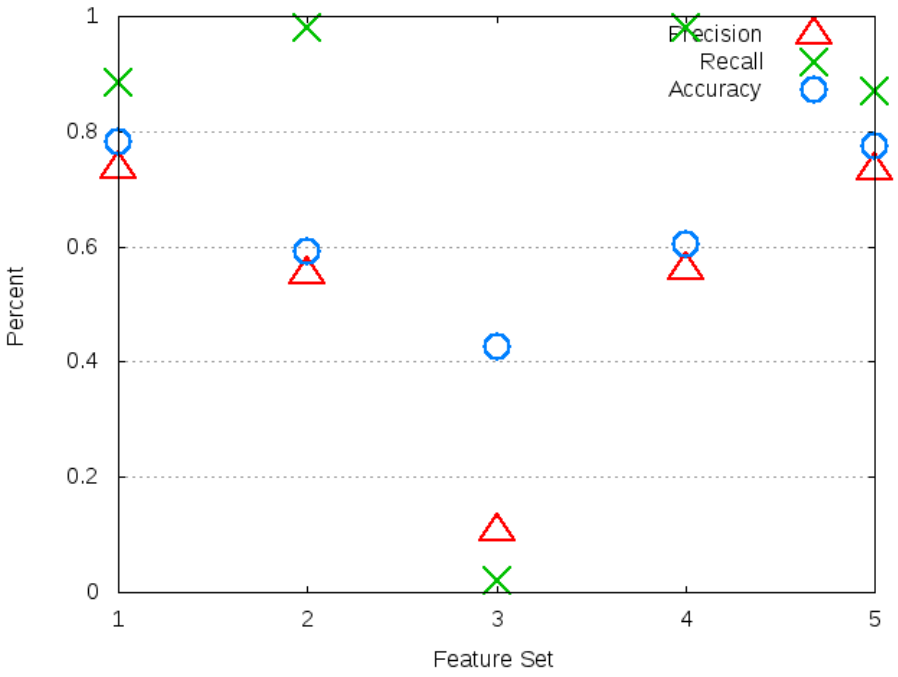


Figure A.87: Feature for ShowcaseView using SVM

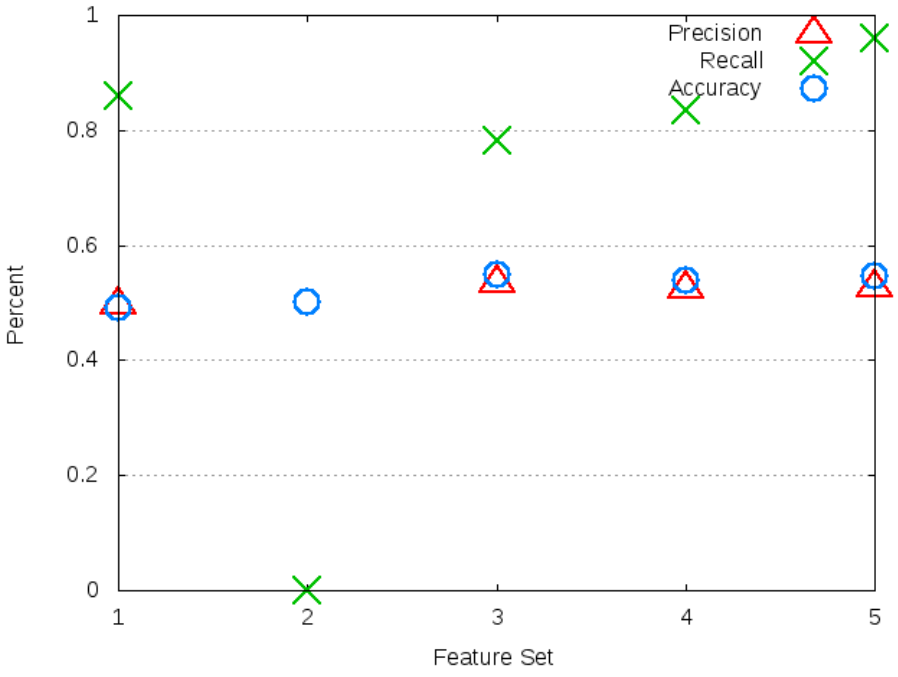


Figure A.88: Feature for smile using SVM

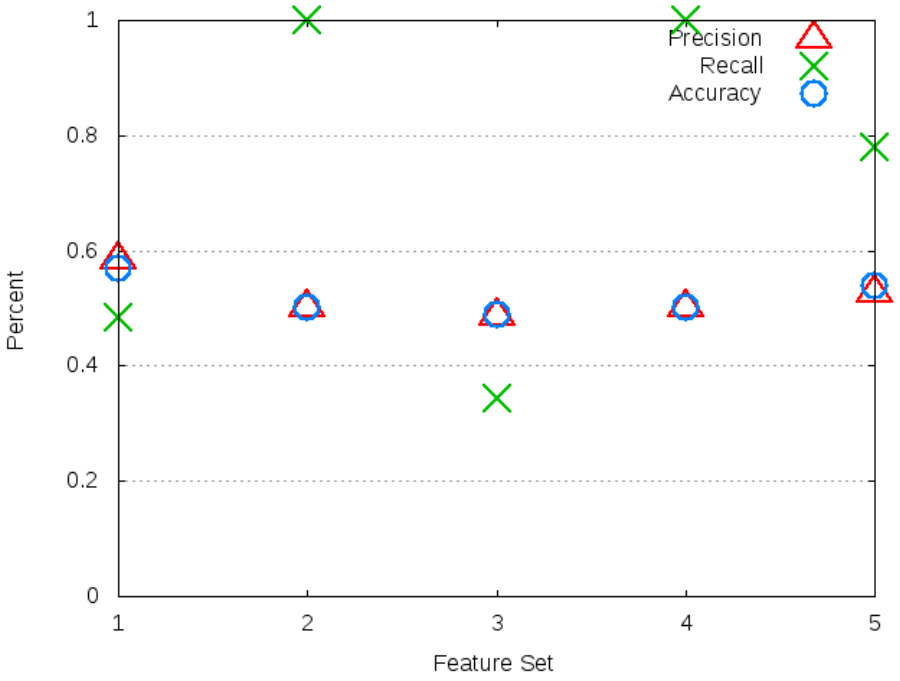


Figure A.89: Feature for spark using SVM

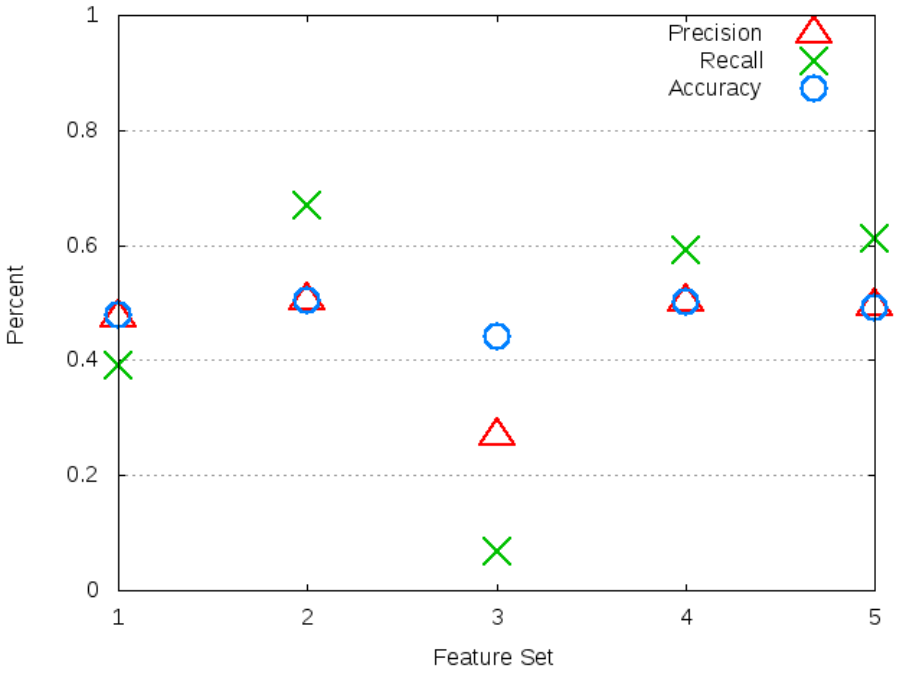


Figure A.90: Feature for storm using SVM

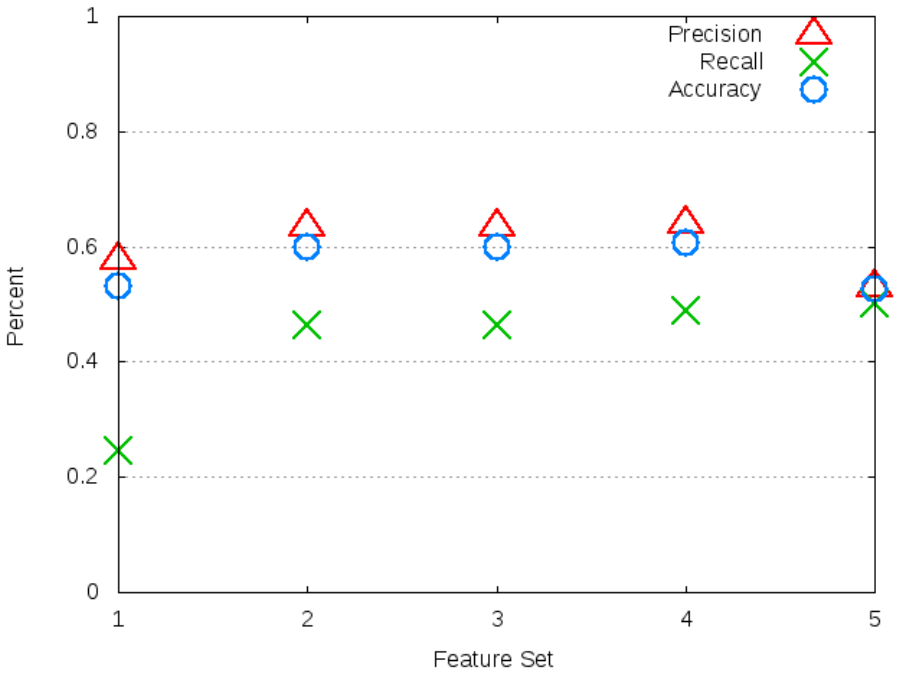


Figure A.91: Feature for tempto using SVM

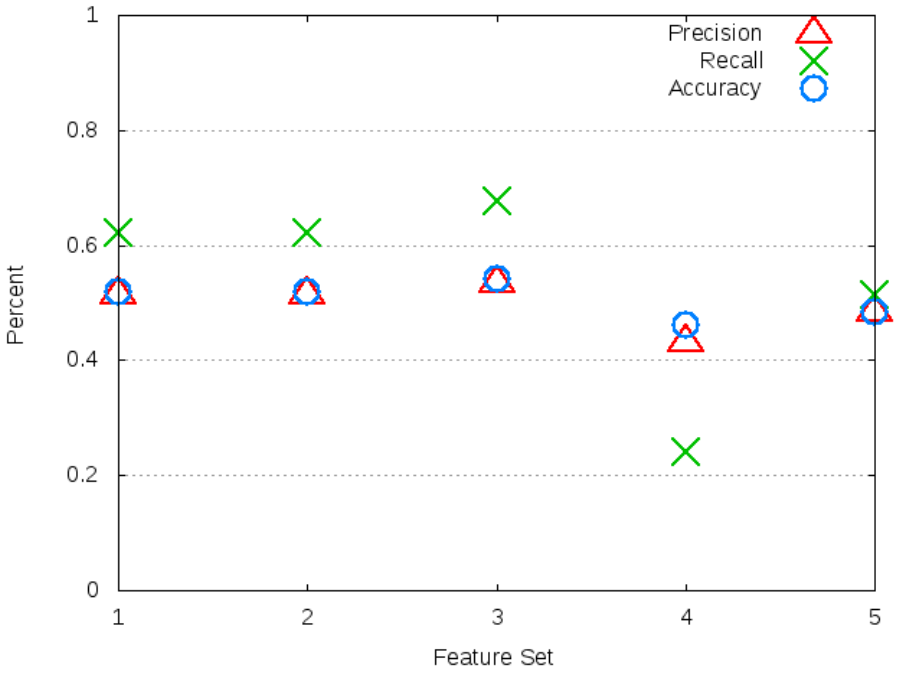


Figure A.92: Feature for yardstick using SVM

A.2.2 Random Forest

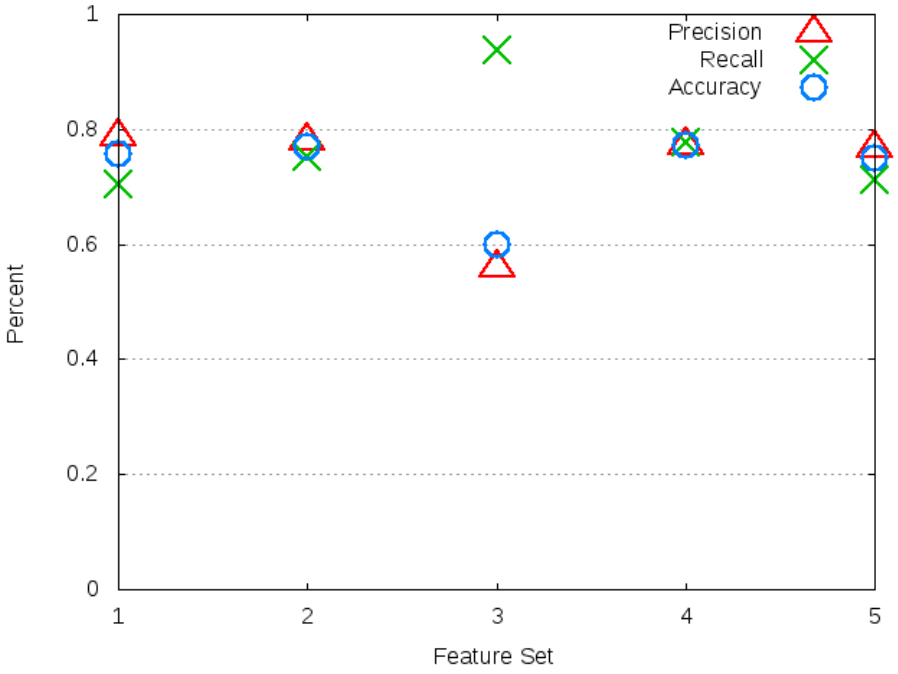


Figure A.93: Feature for acra using RF

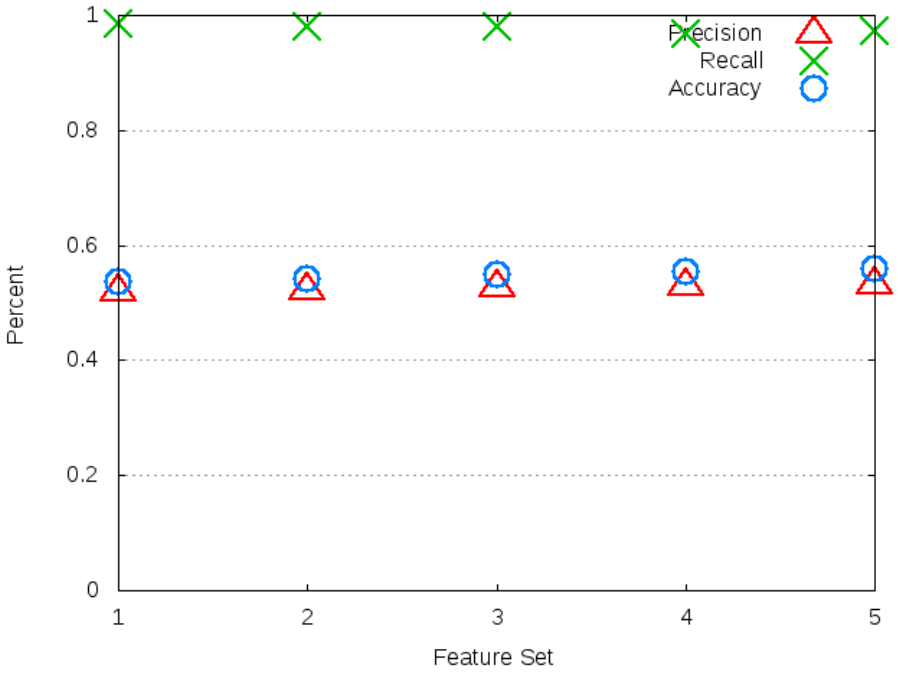


Figure A.94: Feature for arquillian-core using RF

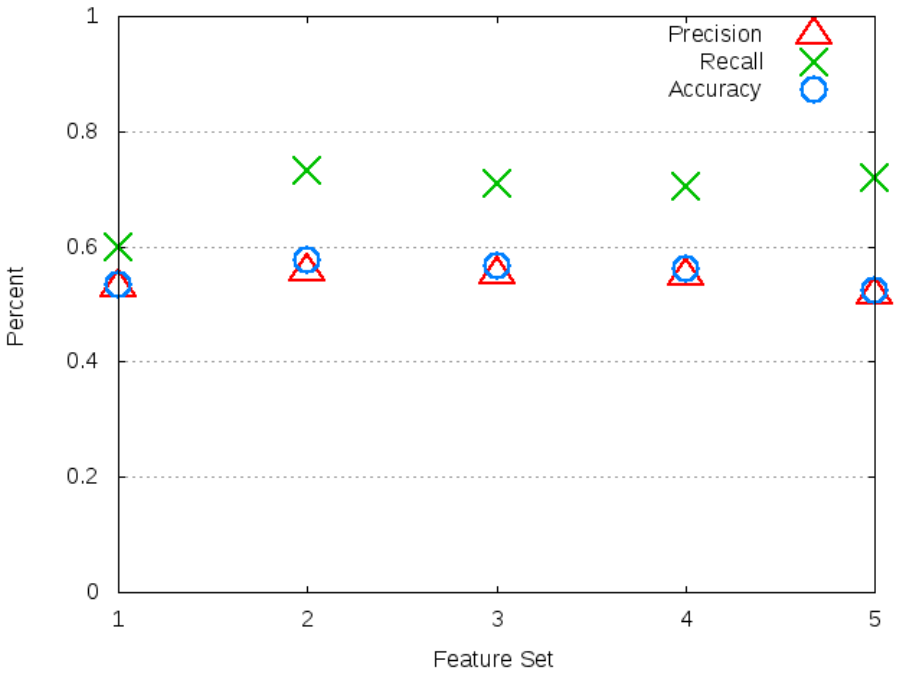


Figure A.95: Feature for blockly-android using RF

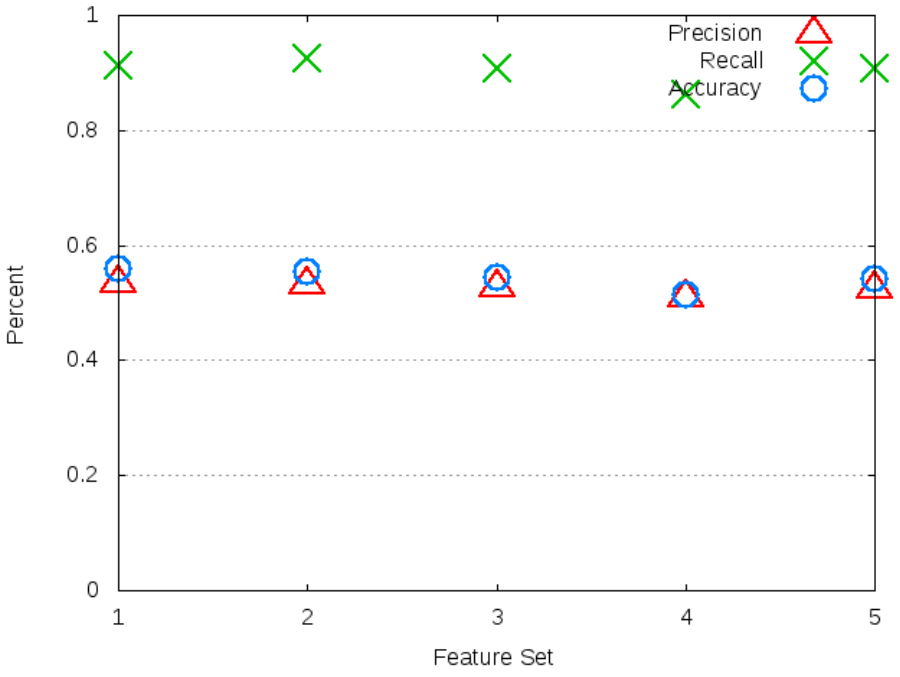


Figure A.96: Feature for brave using RF

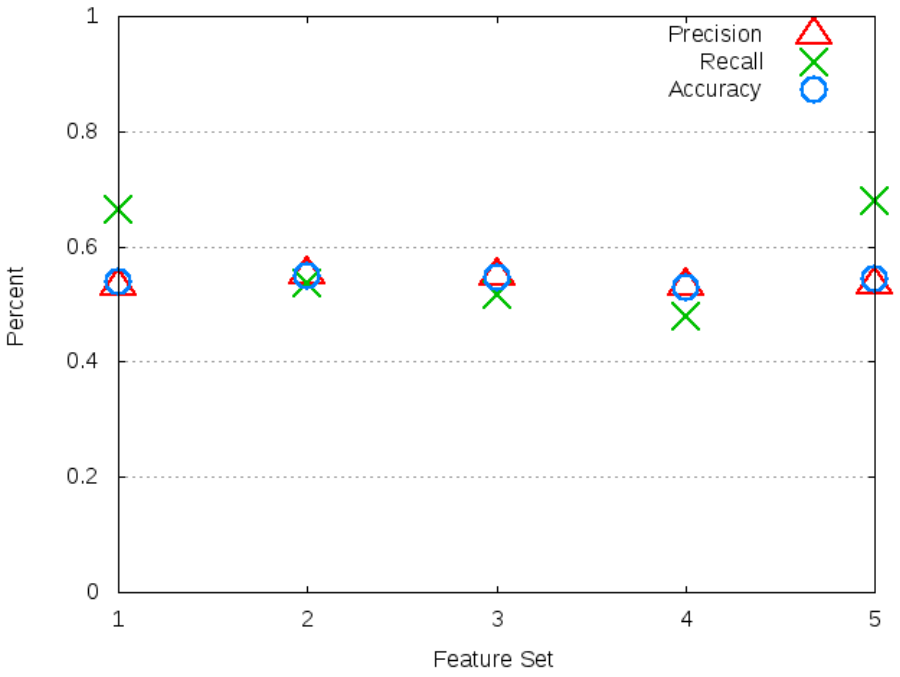


Figure A.97: Feature for cardslib using RF

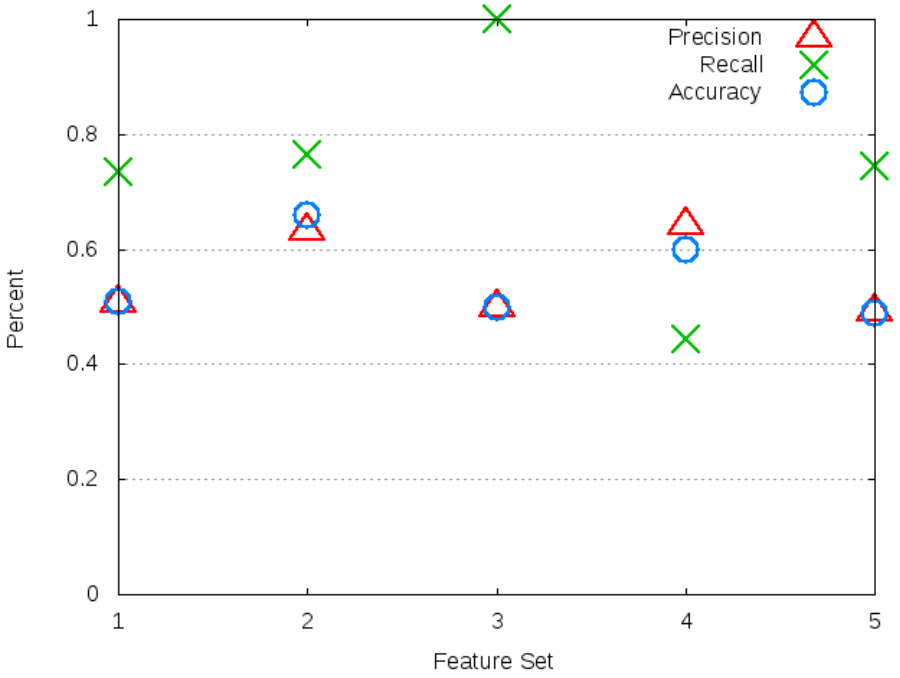


Figure A.98: Feature for dagger using RF

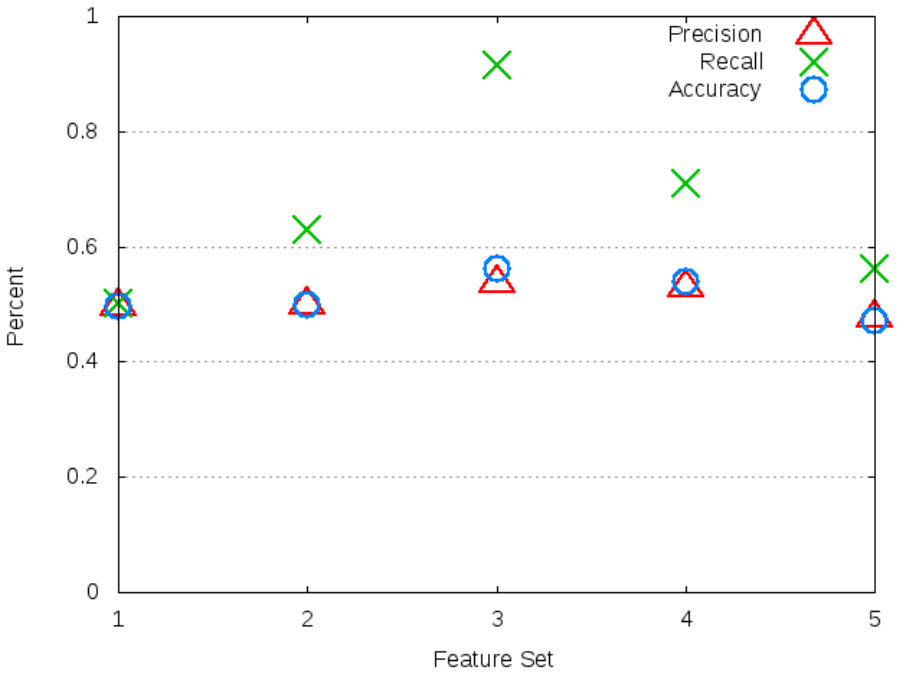


Figure A.99: Feature for deeplearning4j using RF

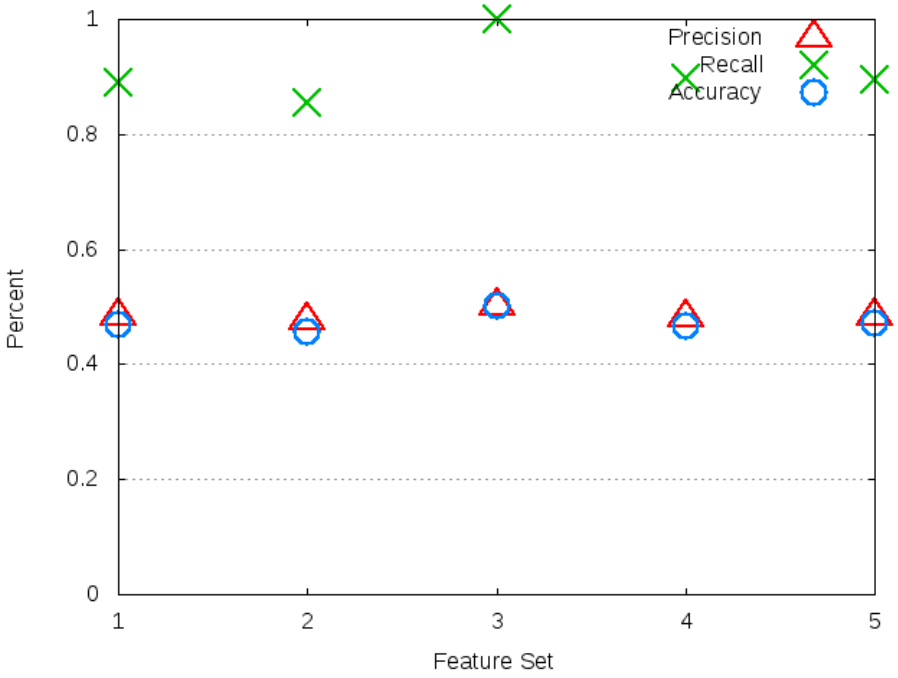


Figure A.100: Feature for fresco using RF

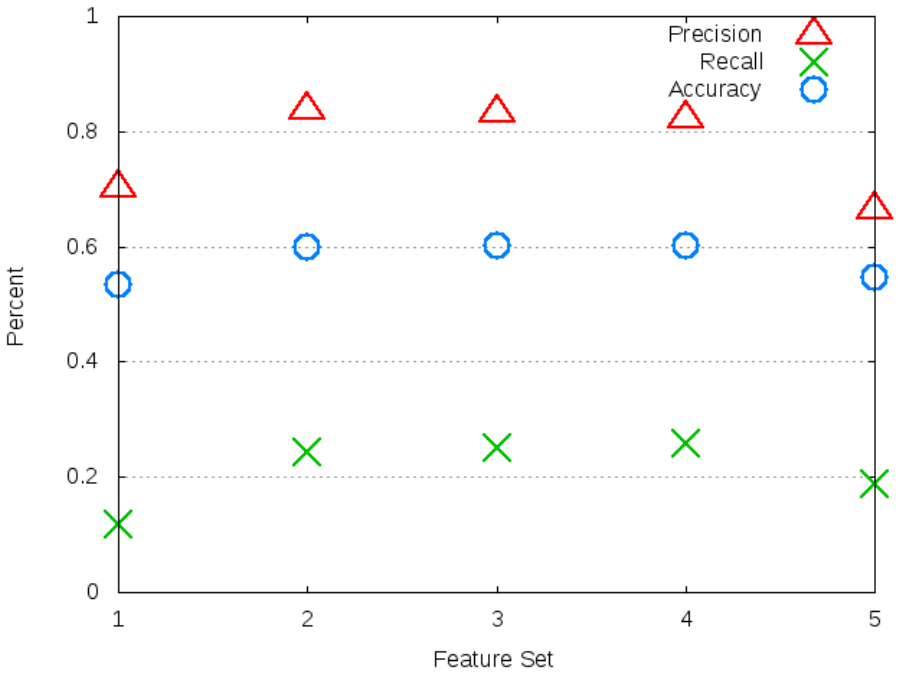


Figure A.101: Feature for governator using RF

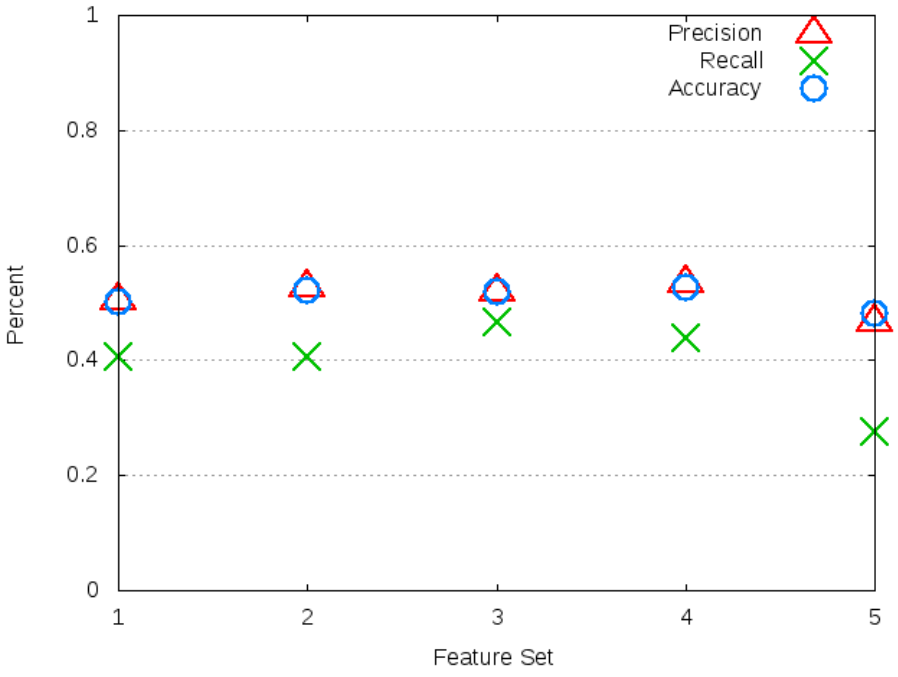


Figure A.102: Feature for greenDAO using RF

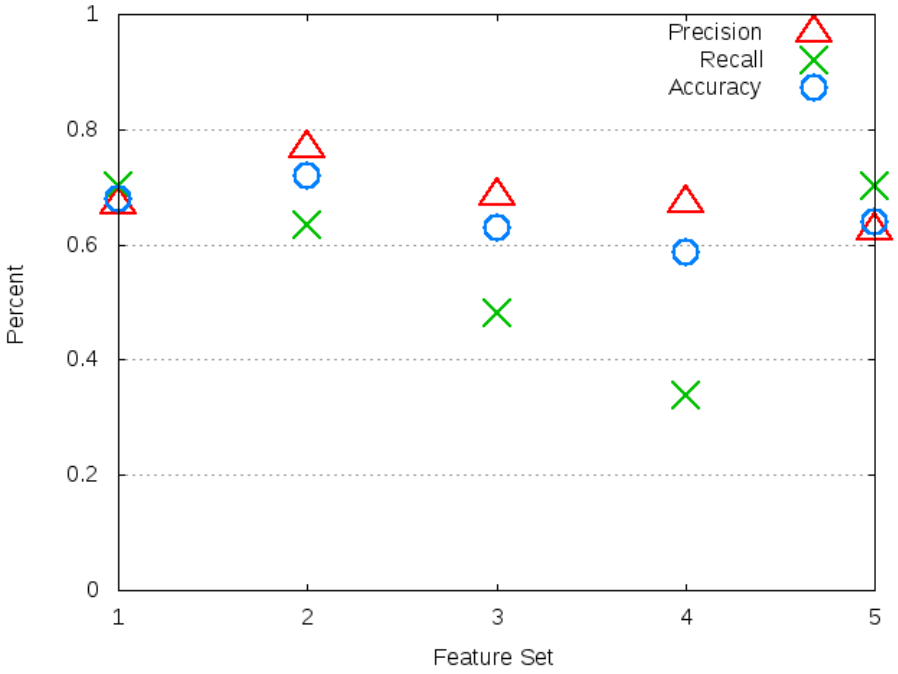


Figure A.103: Feature for http-request using RF

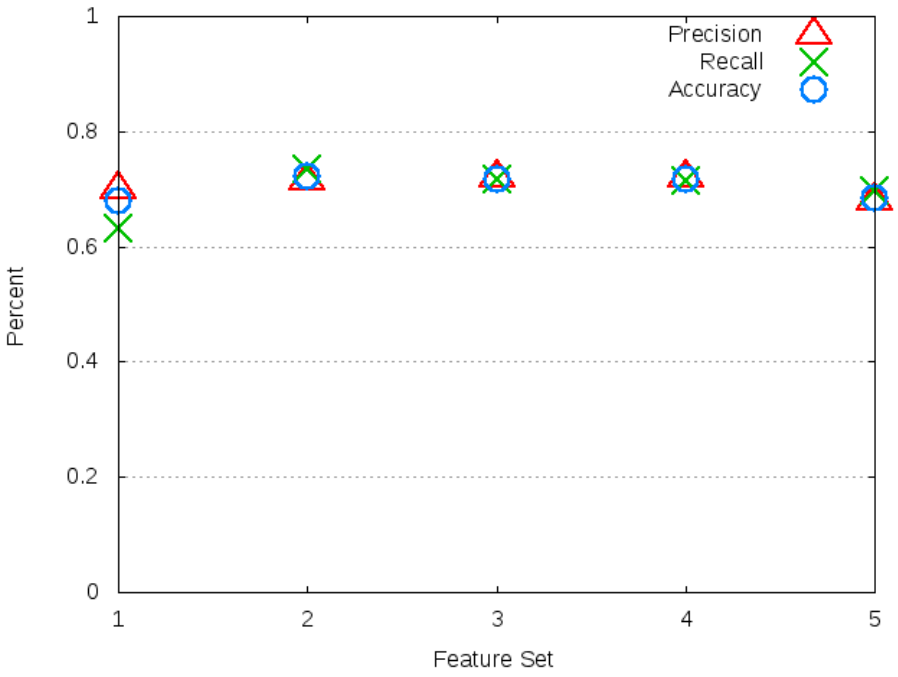


Figure A.104: Feature for ion using RF

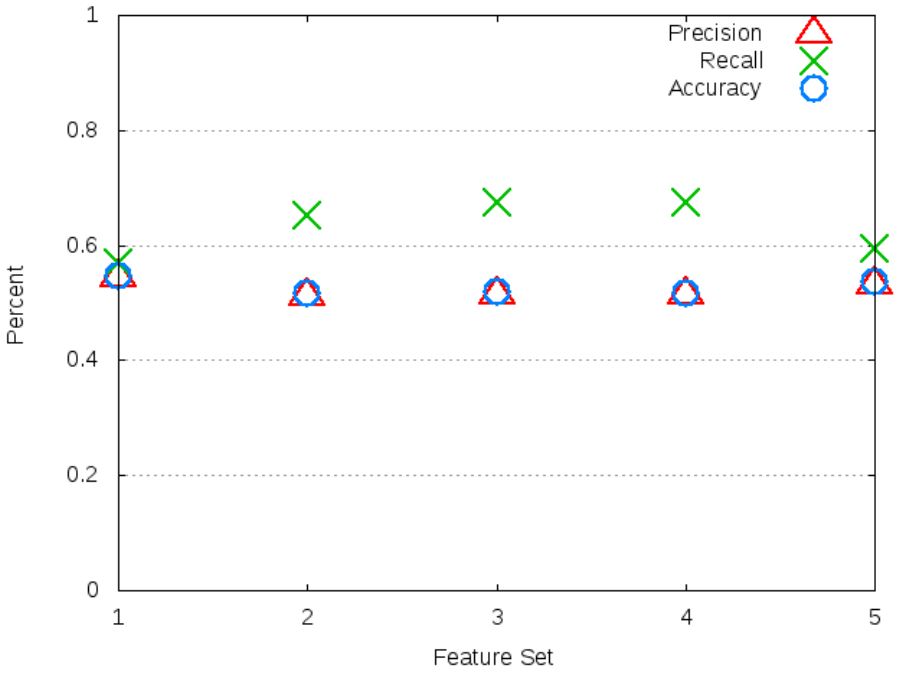


Figure A.105: Feature for jadx using RF

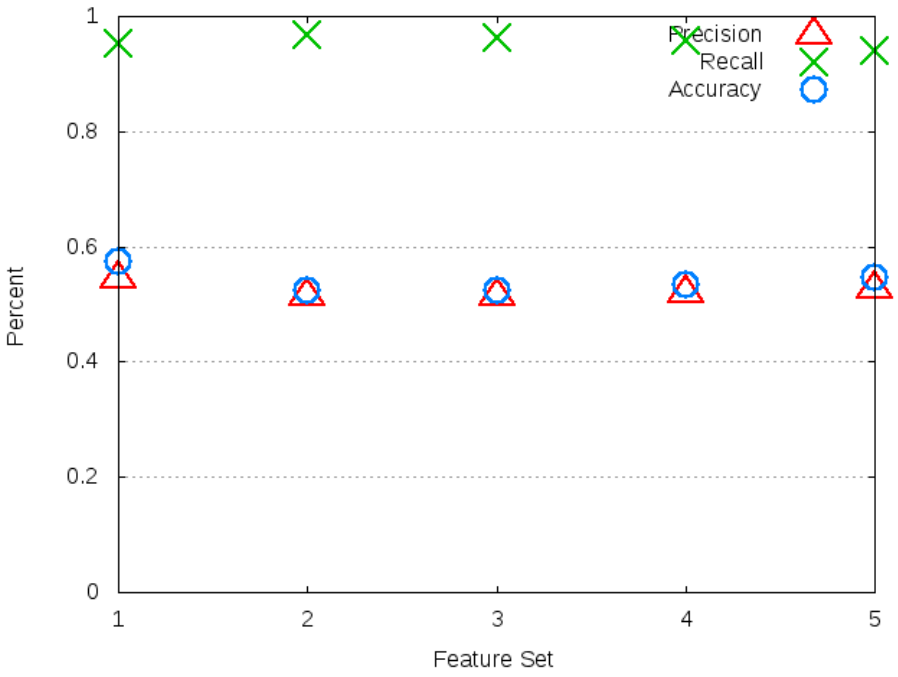


Figure A.106: Feature for mapstruct using RF

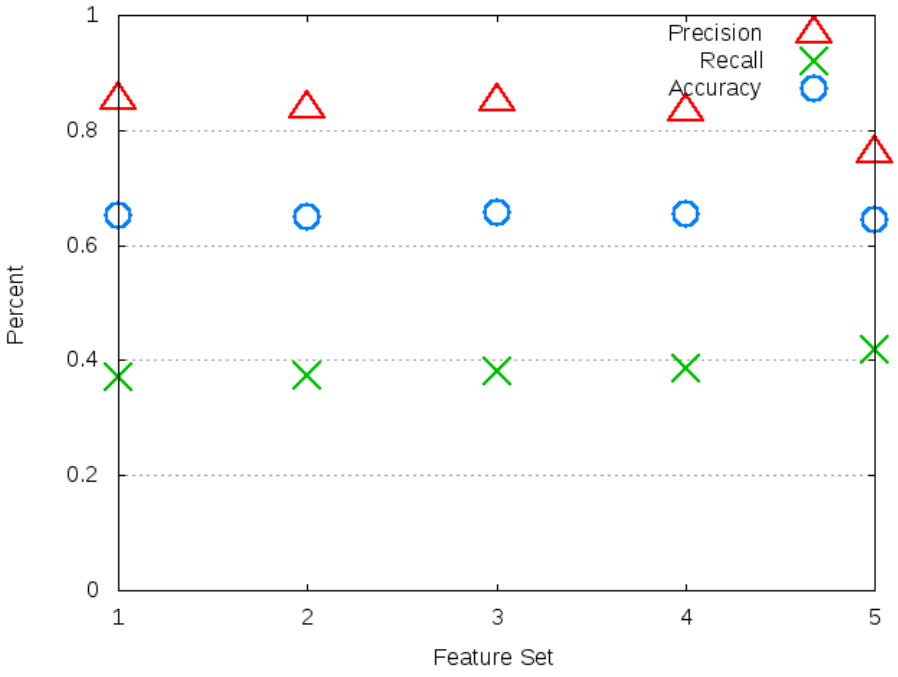


Figure A.107: Feature for nettosphere using RF

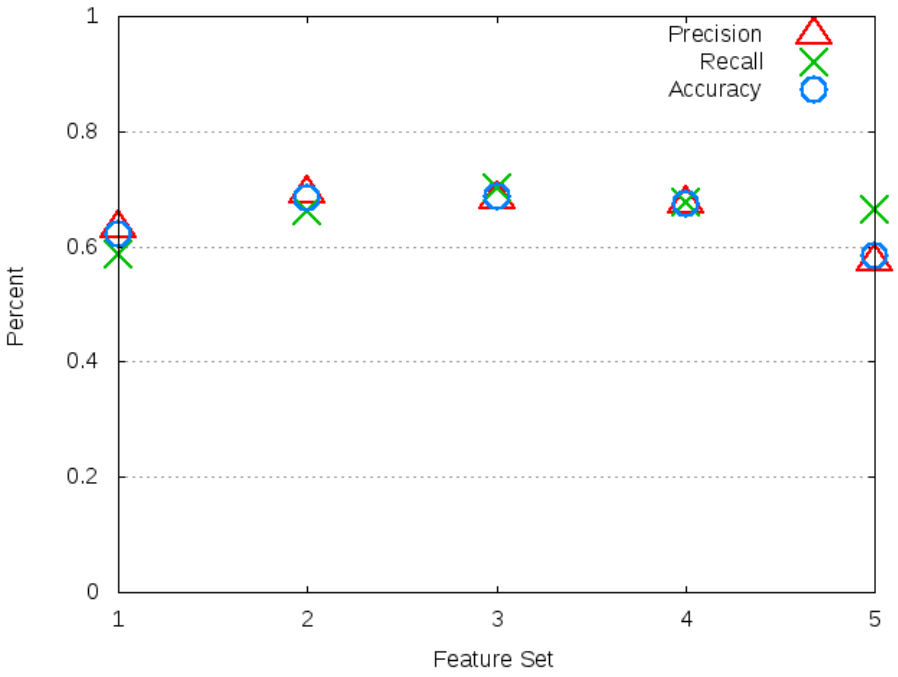


Figure A.108: Feature for parceler using RF

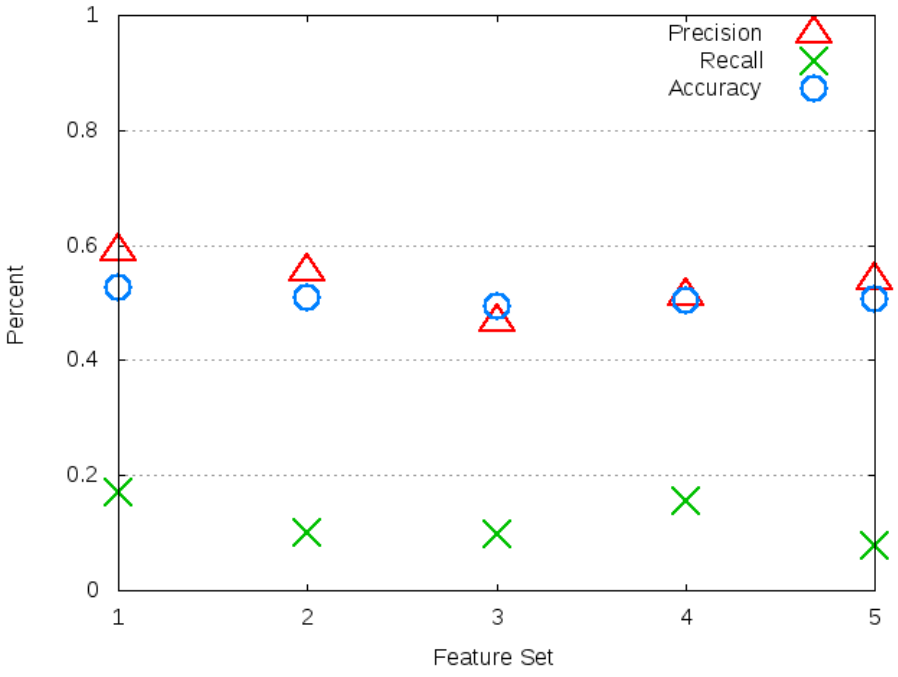


Figure A.109: Feature for retrolambda using RF

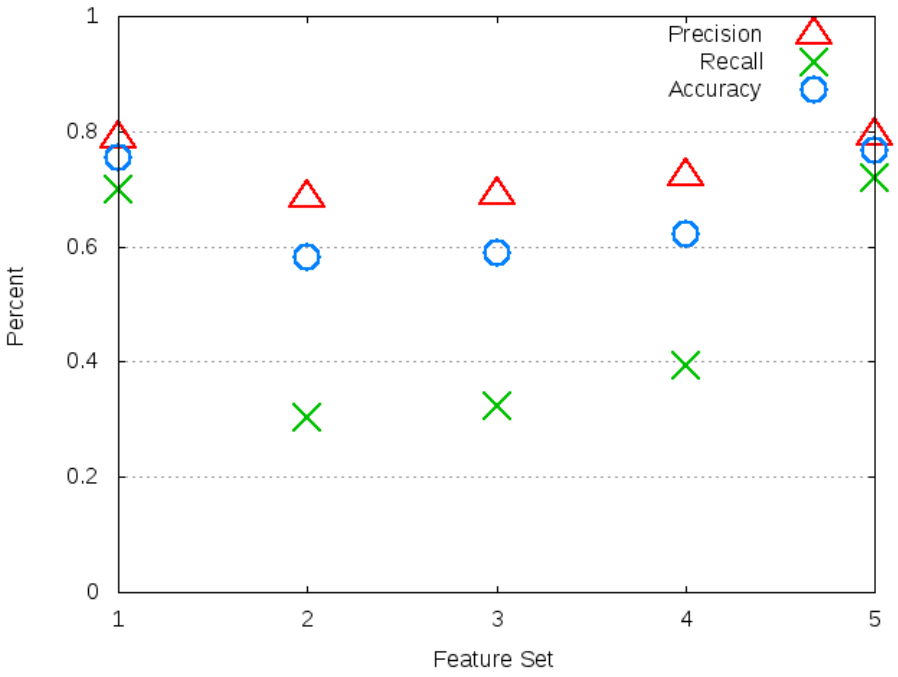


Figure A.110: Feature for ShowcaseView using RF

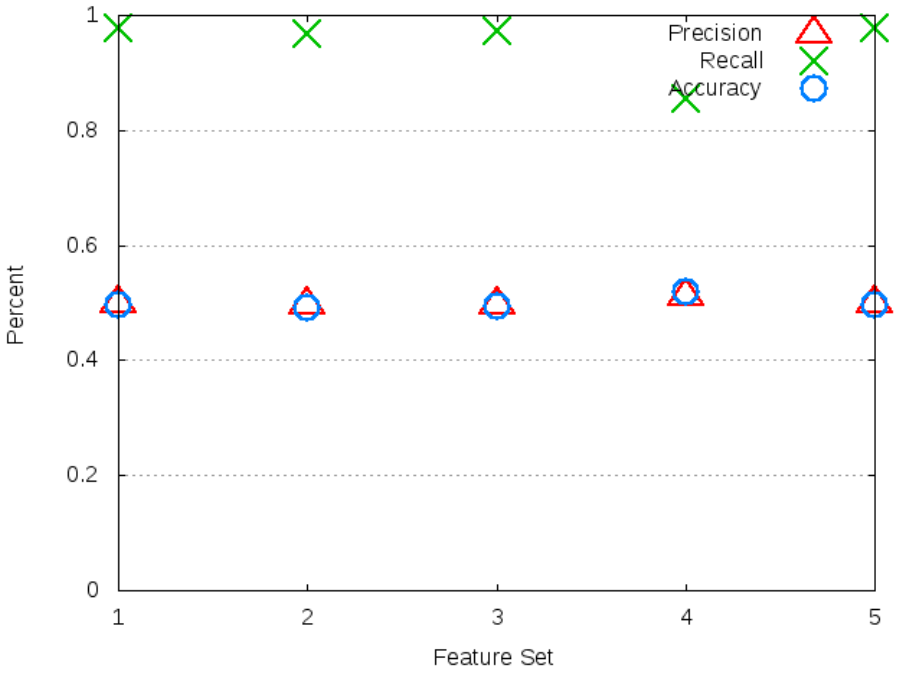


Figure A.111: Feature for smile using RF

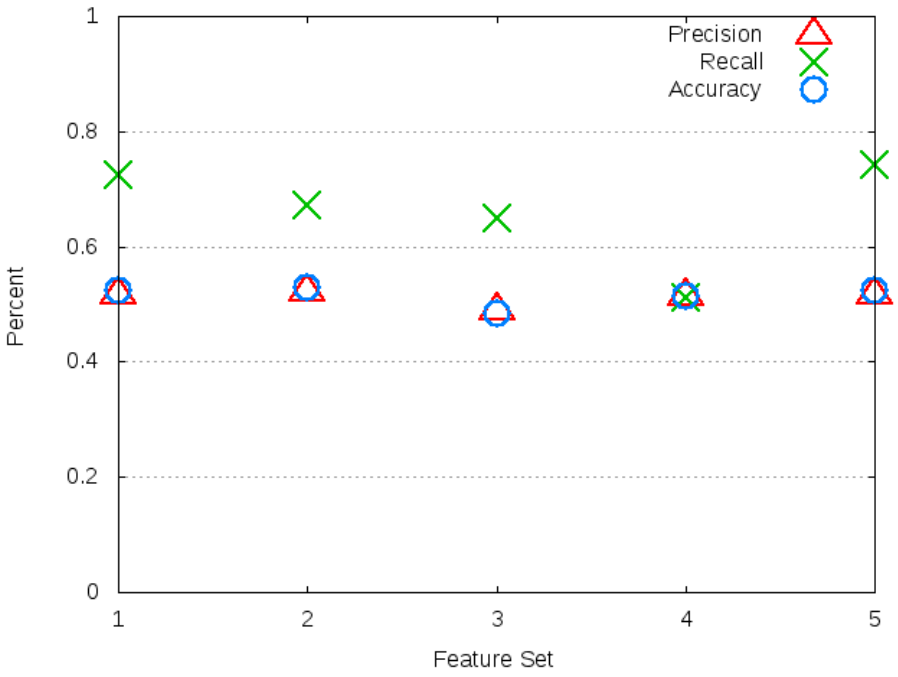


Figure A.112: Feature for spark using RF

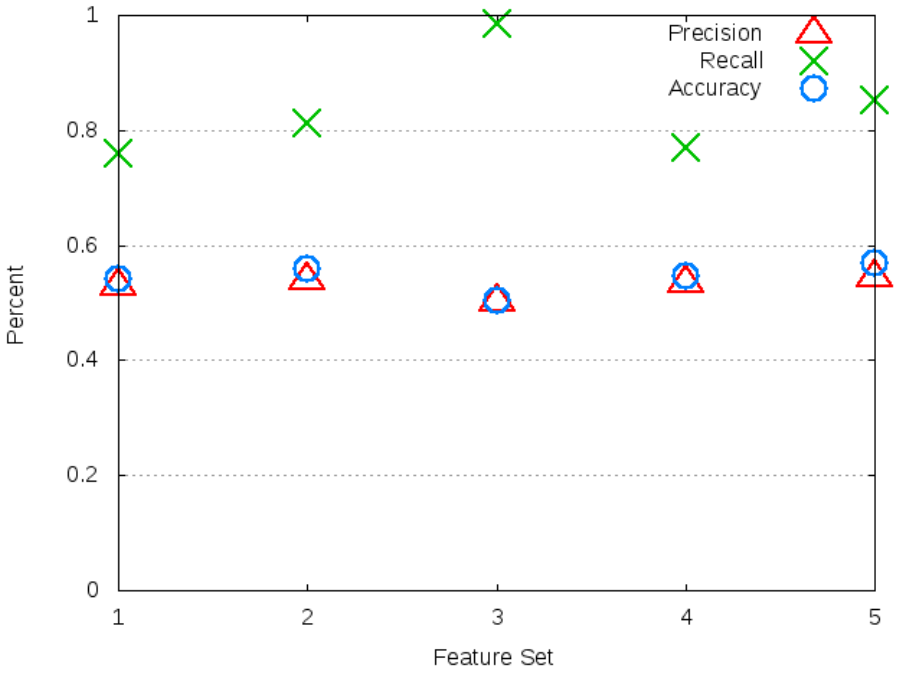


Figure A.113: Feature for storm using RF

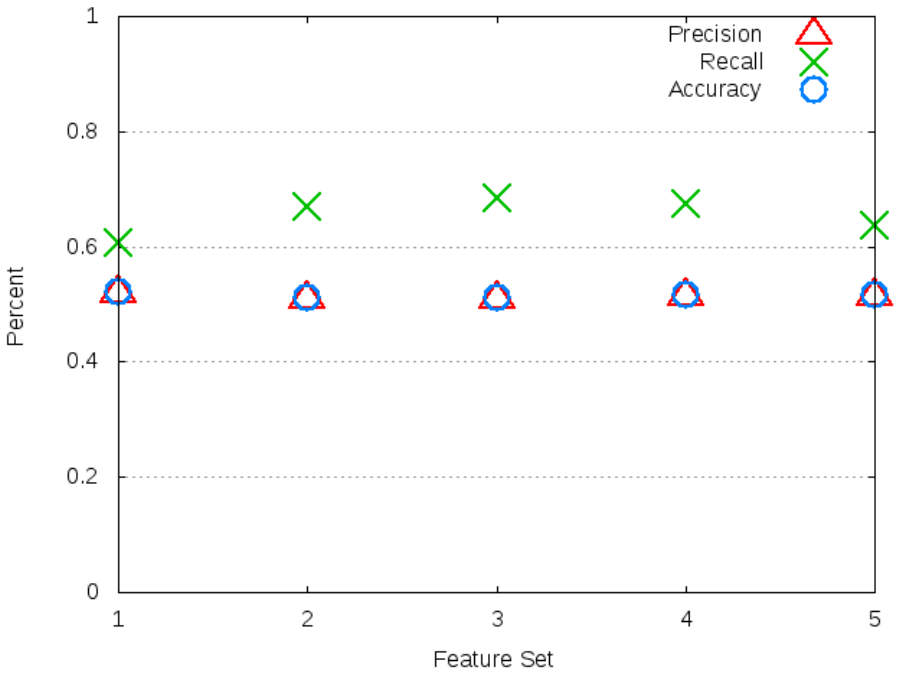


Figure A.114: Feature for tempto using RF

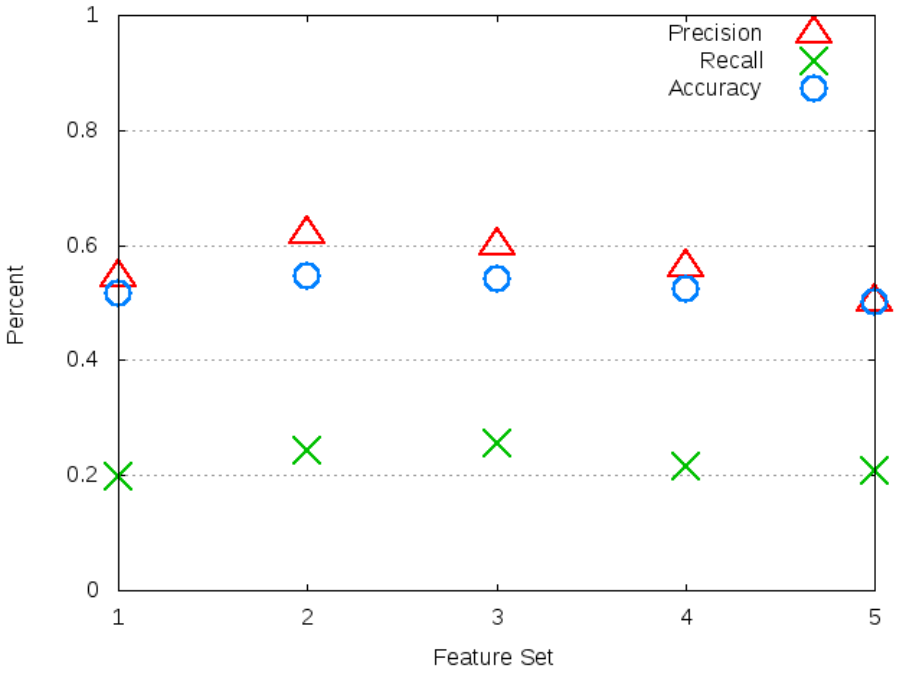


Figure A.115: Feature for yardstick using RF

A.3 Experiment 3

A.3.1 Support Vector Machine

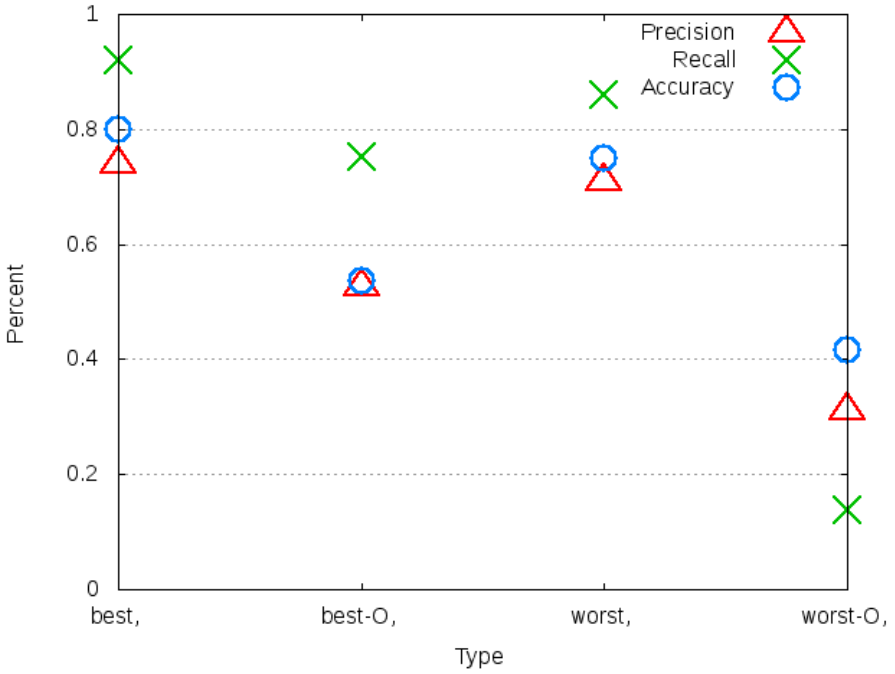


Figure A.116: Oversampling for acra using SVM

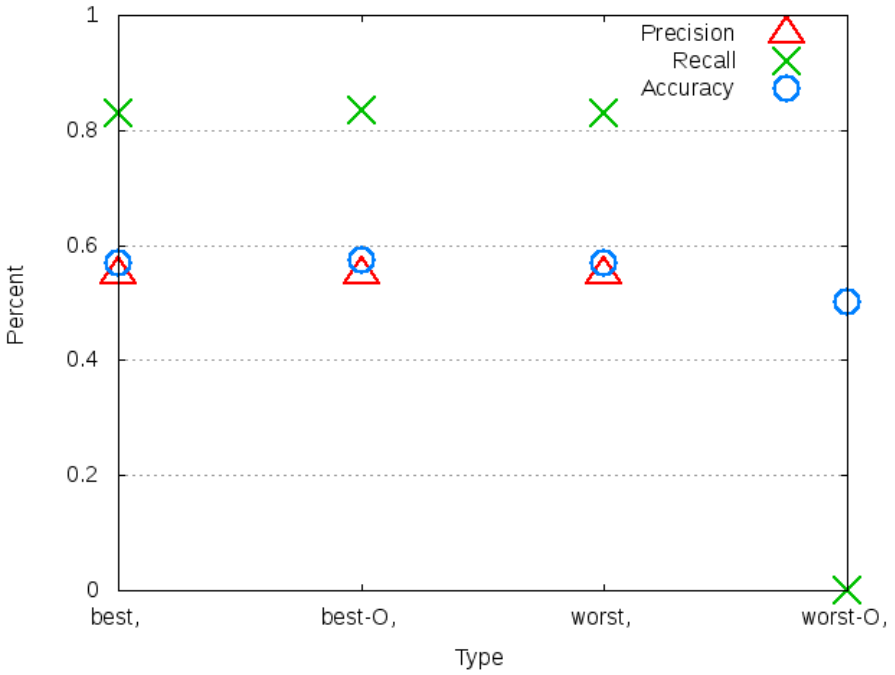


Figure A.117: Oversampling for arquillian-core using SVM

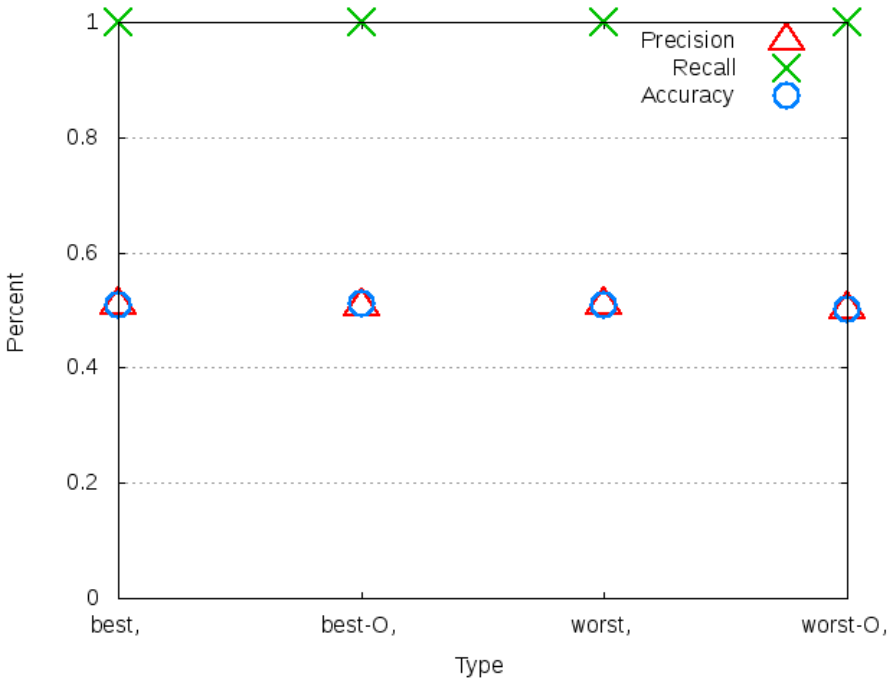


Figure A.118: Oversampling for blockly-android using SVM

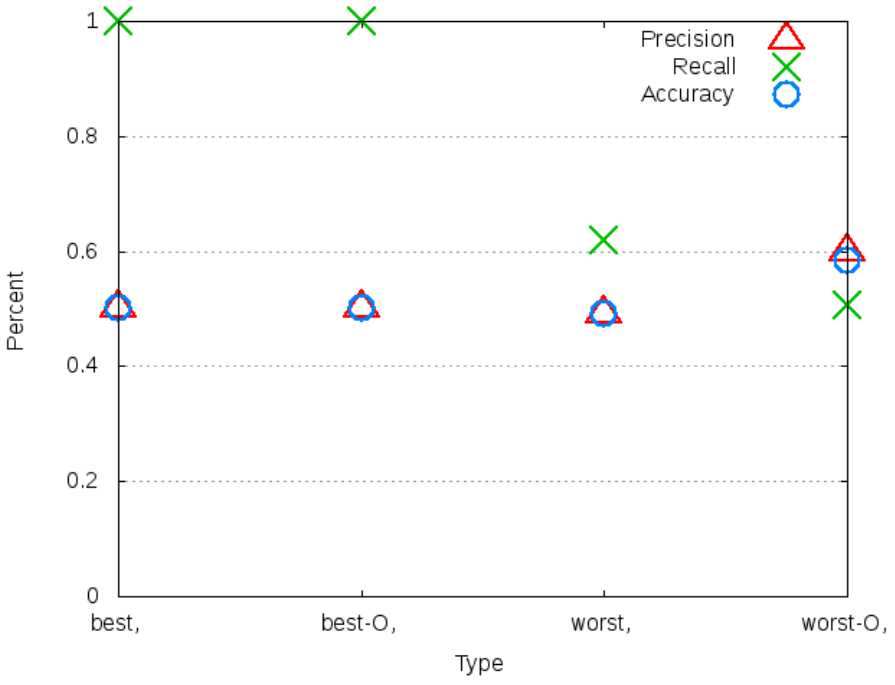


Figure A.119: Oversampling for brave using SVM

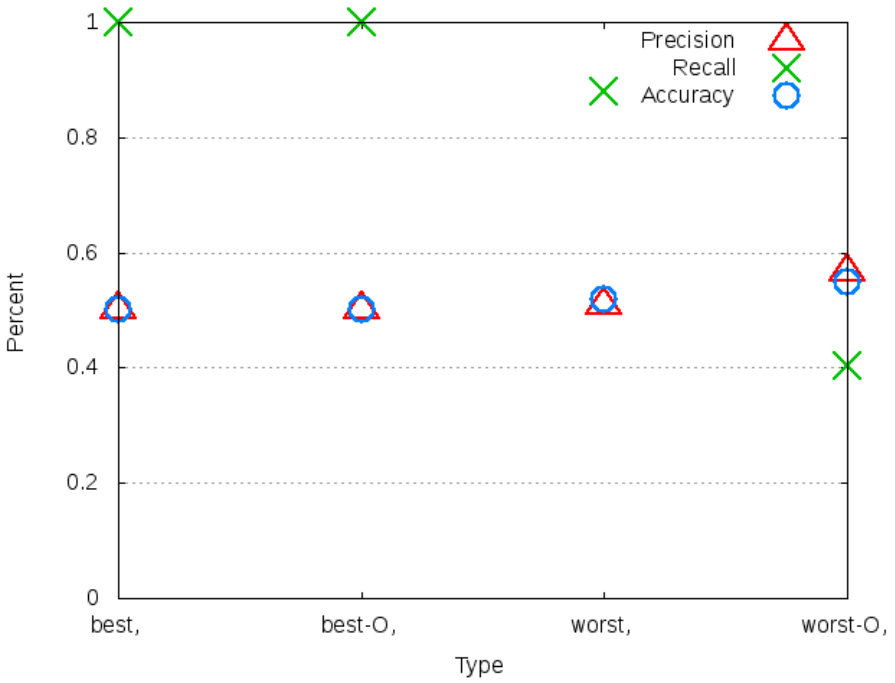


Figure A.120: Oversampling for cardslib using SVM

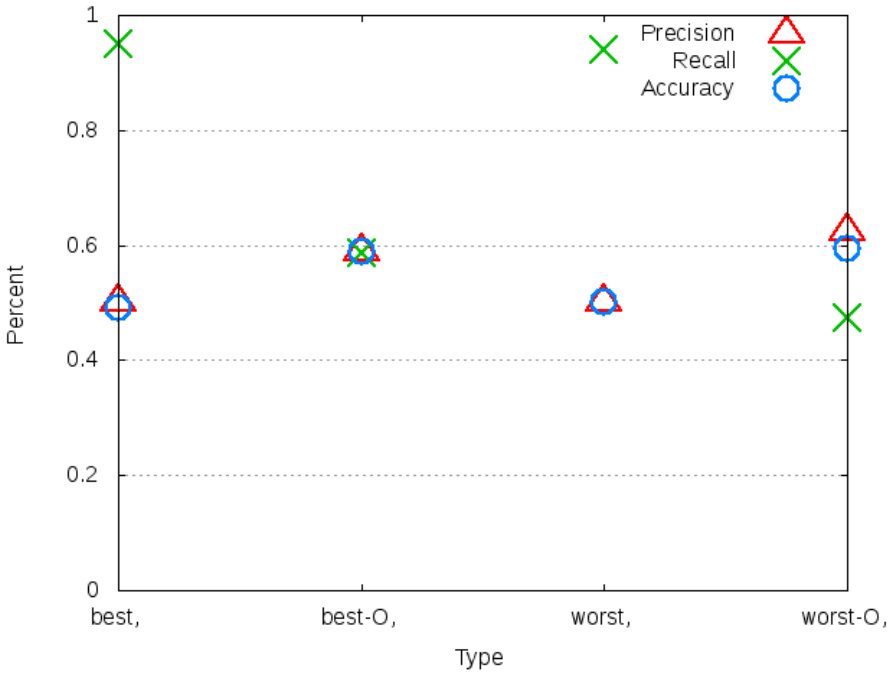


Figure A.121: Oversampling for dagger using SVM

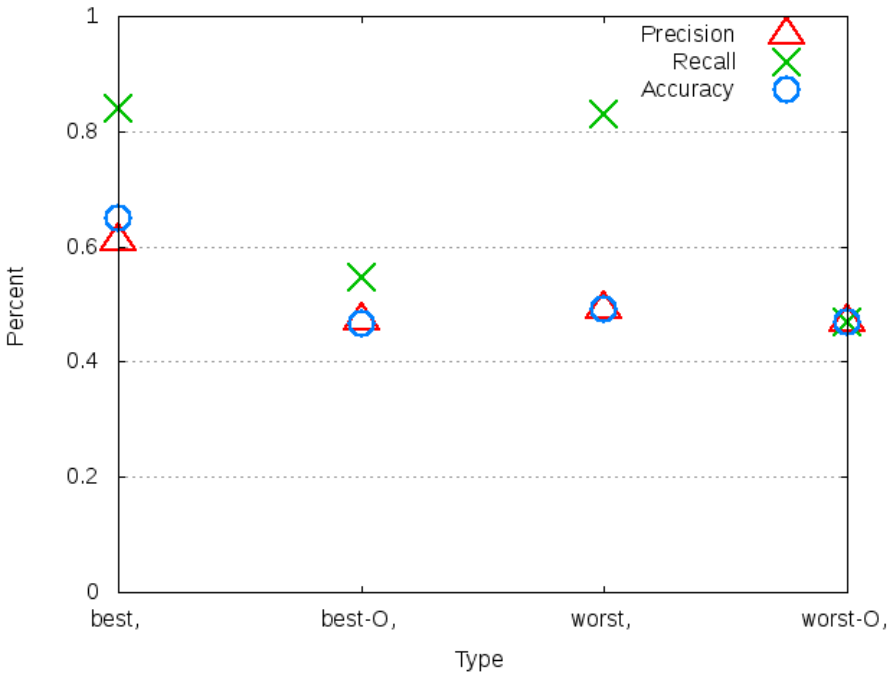


Figure A.122: Oversampling for deeplearning4j using SVM

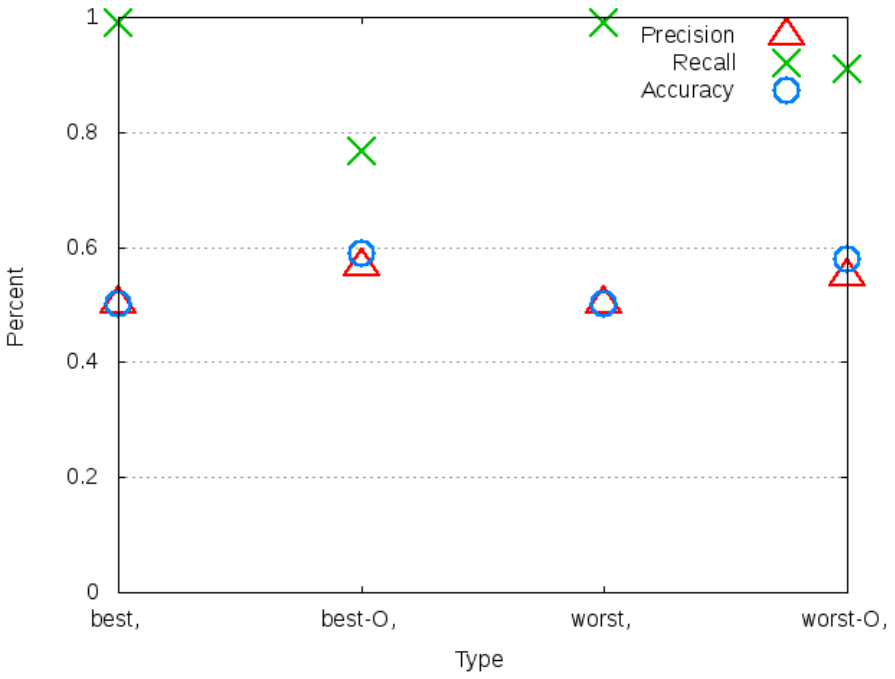


Figure A.123: Oversampling for fresco using SVM

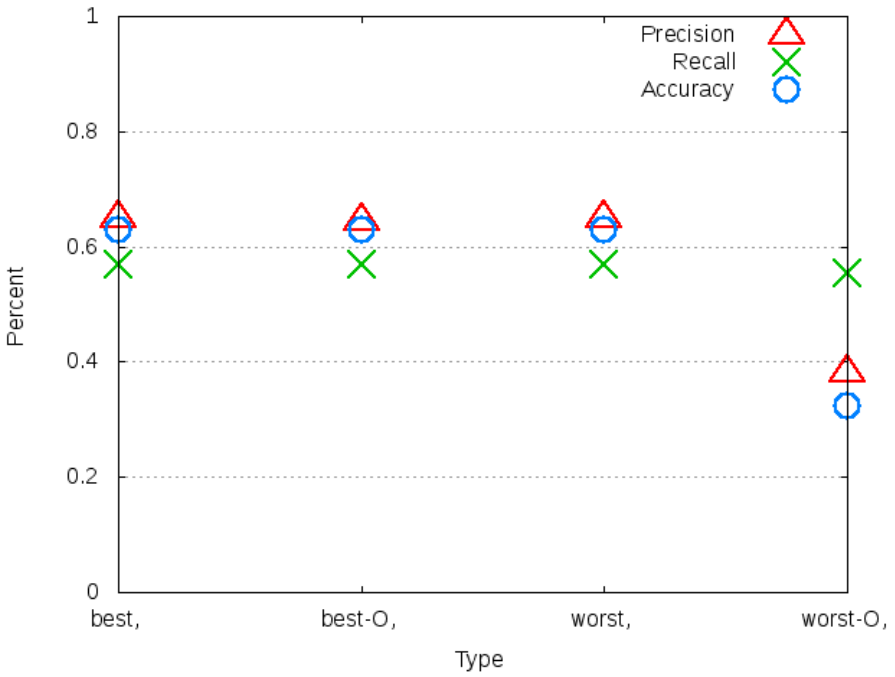


Figure A.124: Oversampling for governator using SVM

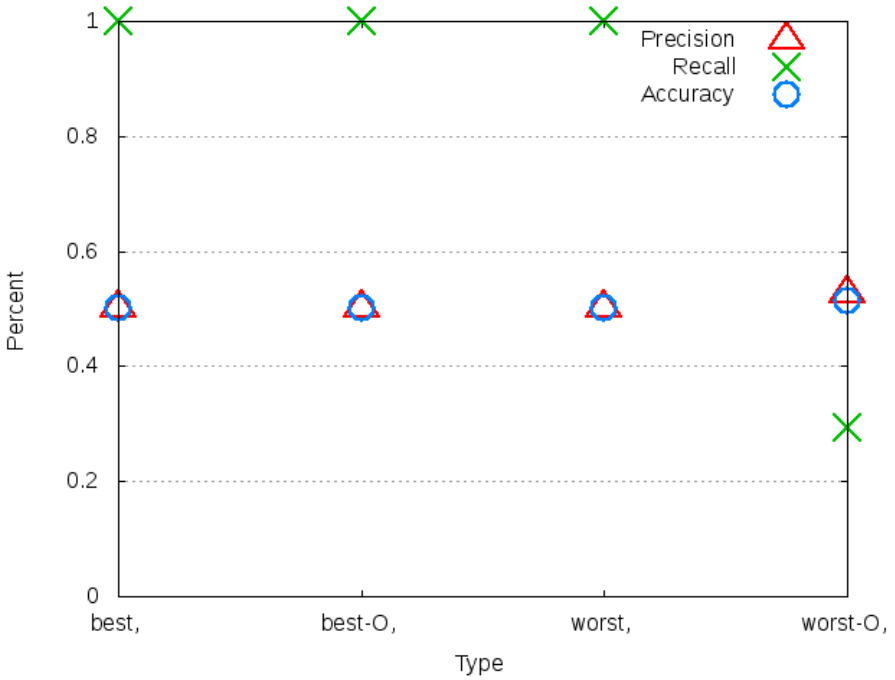


Figure A.125: Oversampling for greenDAO using SVM

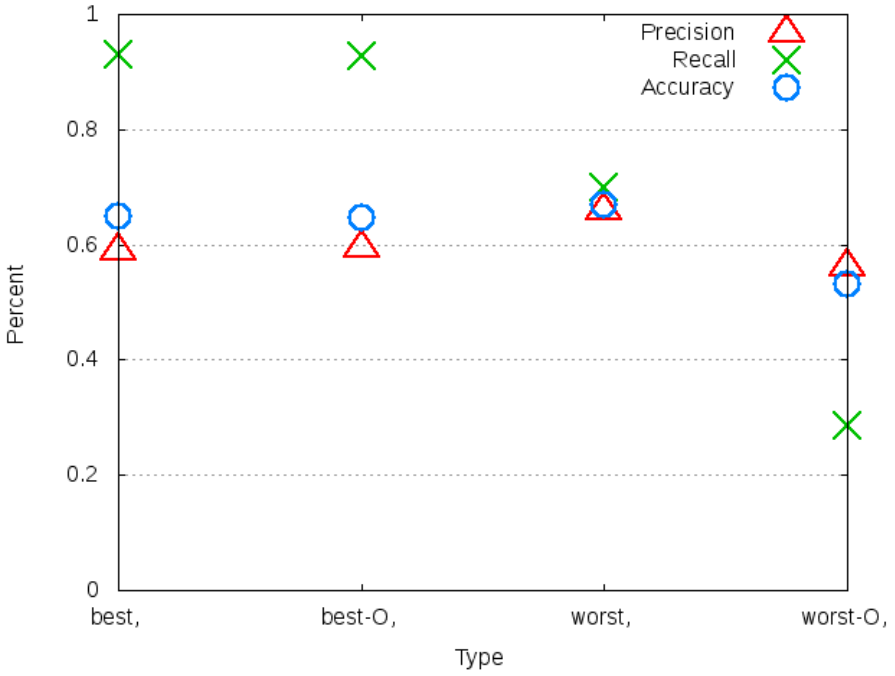


Figure A.126: Oversampling for http-request using SVM

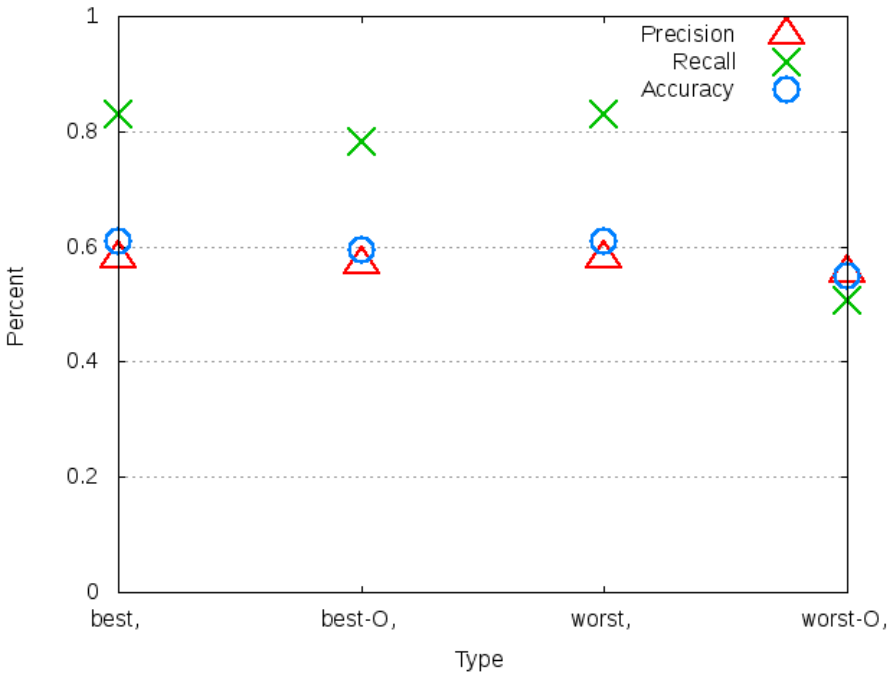


Figure A.127: Oversampling for ion using SVM

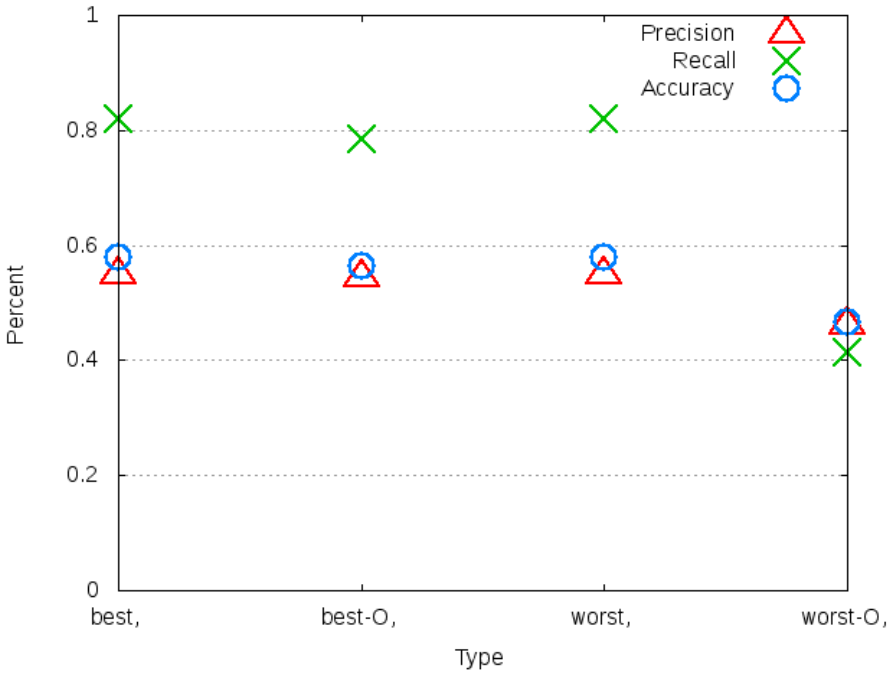


Figure A.128: Oversampling for jadx using SVM

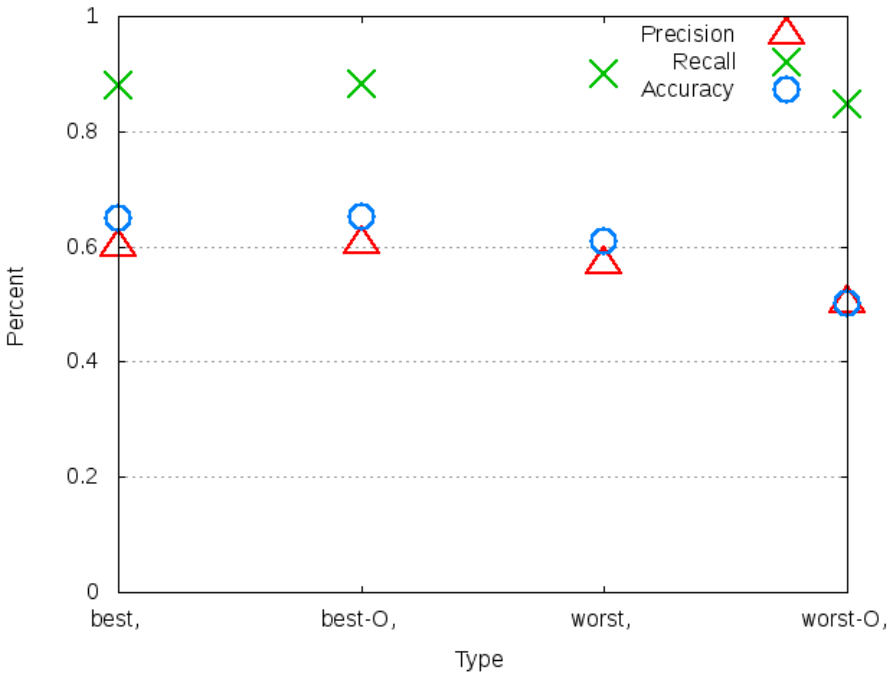


Figure A.129: Oversampling for mapstruct using SVM

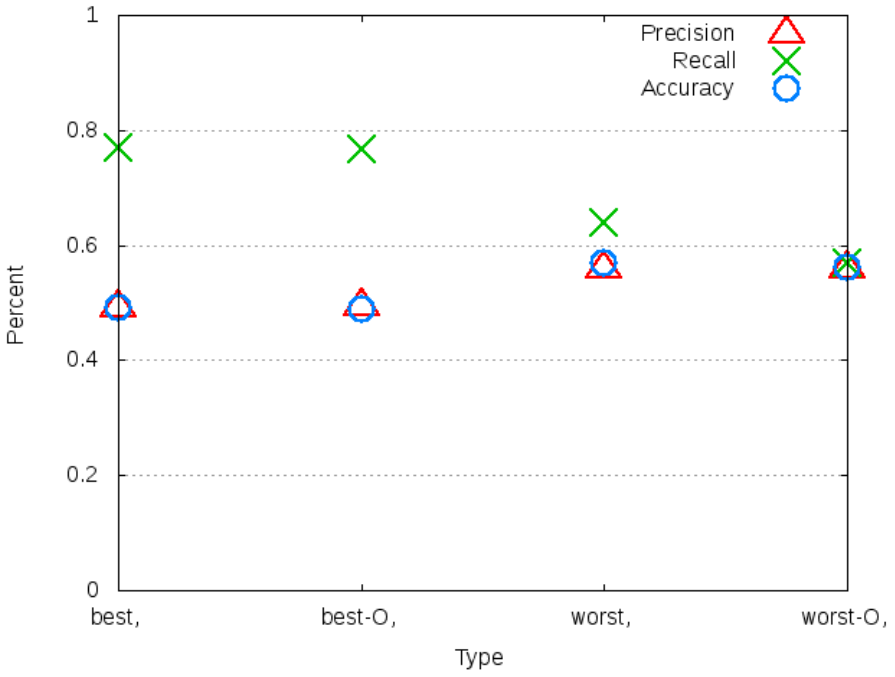


Figure A.130: Oversampling for nettosphere using SVM

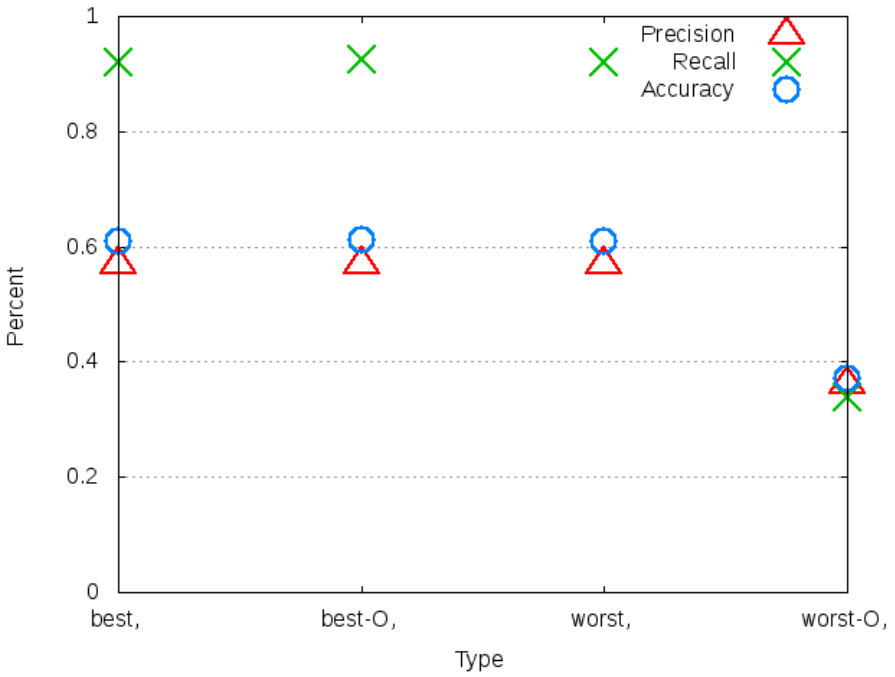


Figure A.131: Oversampling for parceler using SVM

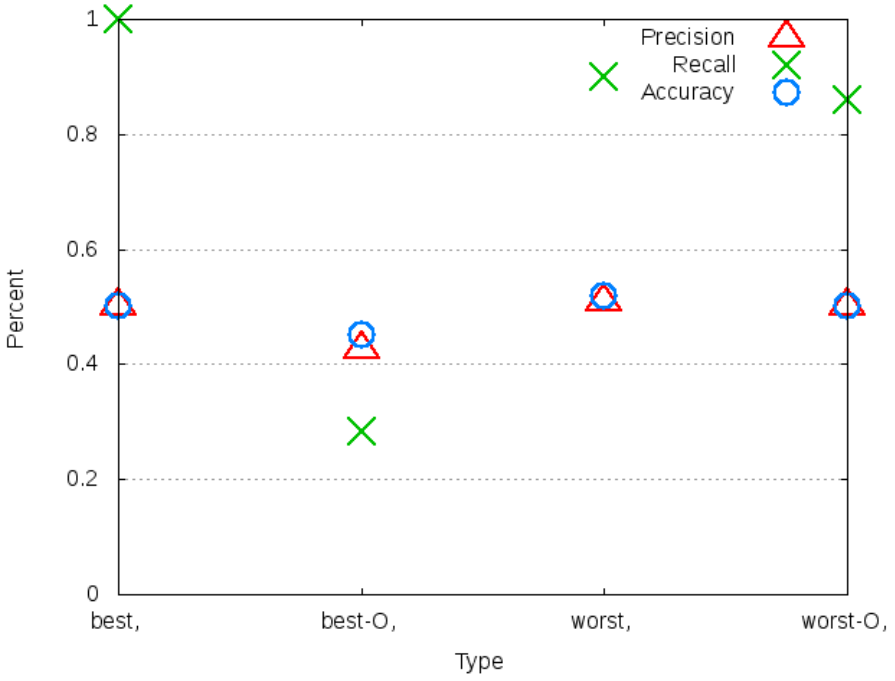


Figure A.132: Oversampling for retrolambda using SVM

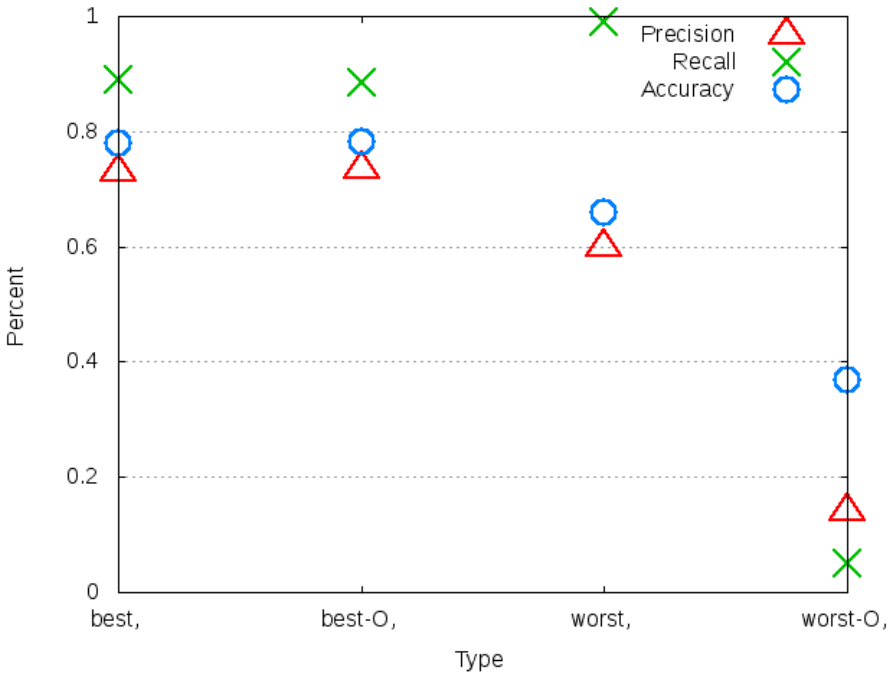


Figure A.133: Oversampling for ShowcaseView using SVM

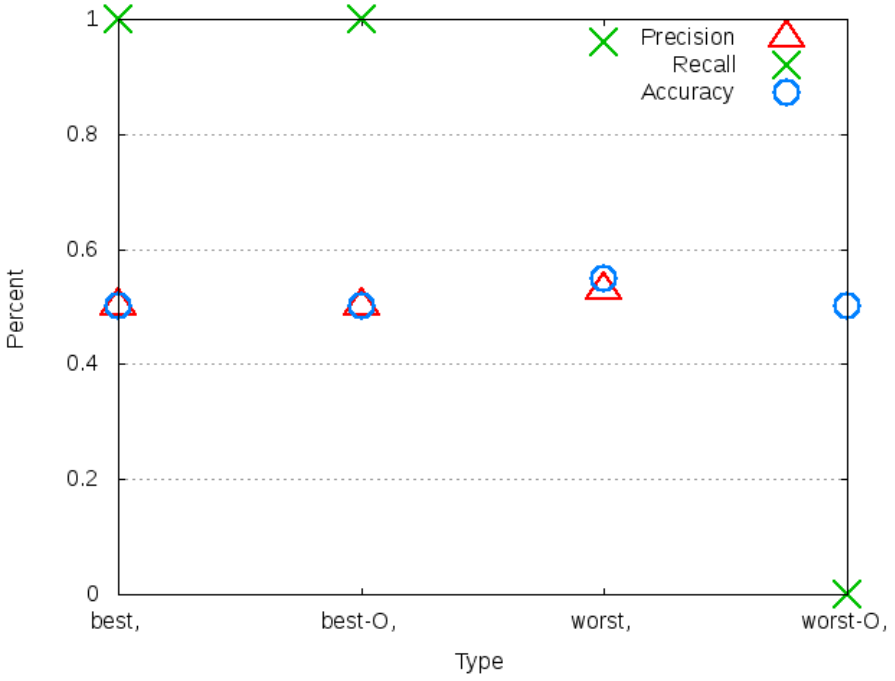


Figure A.134: Oversampling for smile using SVM

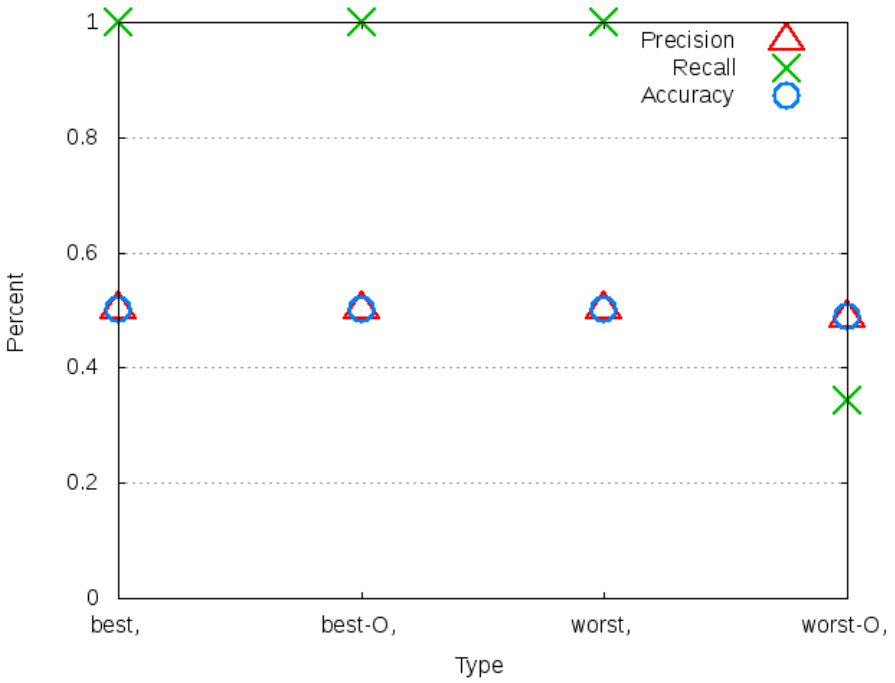


Figure A.135: Oversampling for spark using SVM

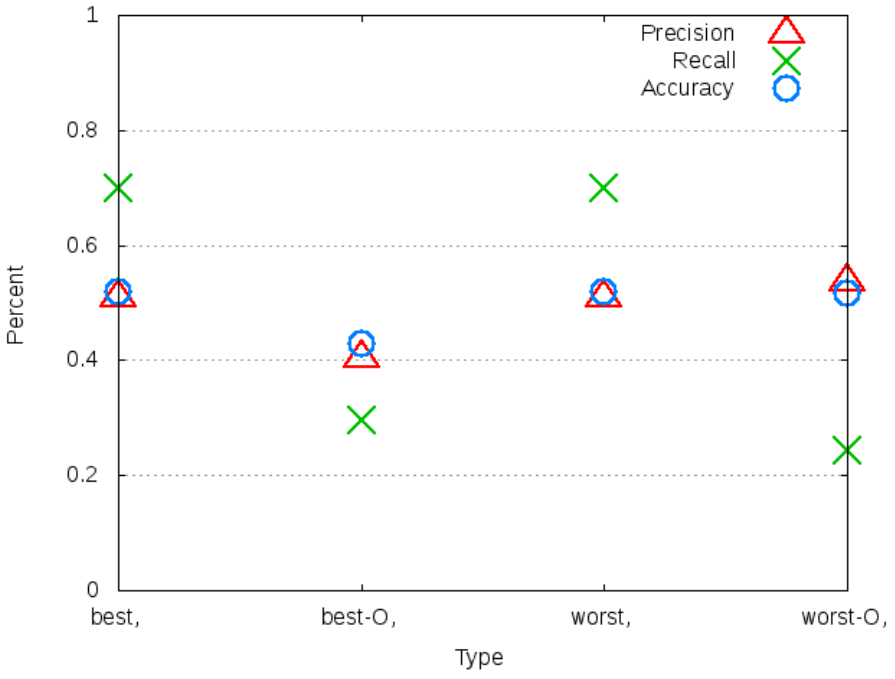


Figure A.136: Oversampling for storm using SVM

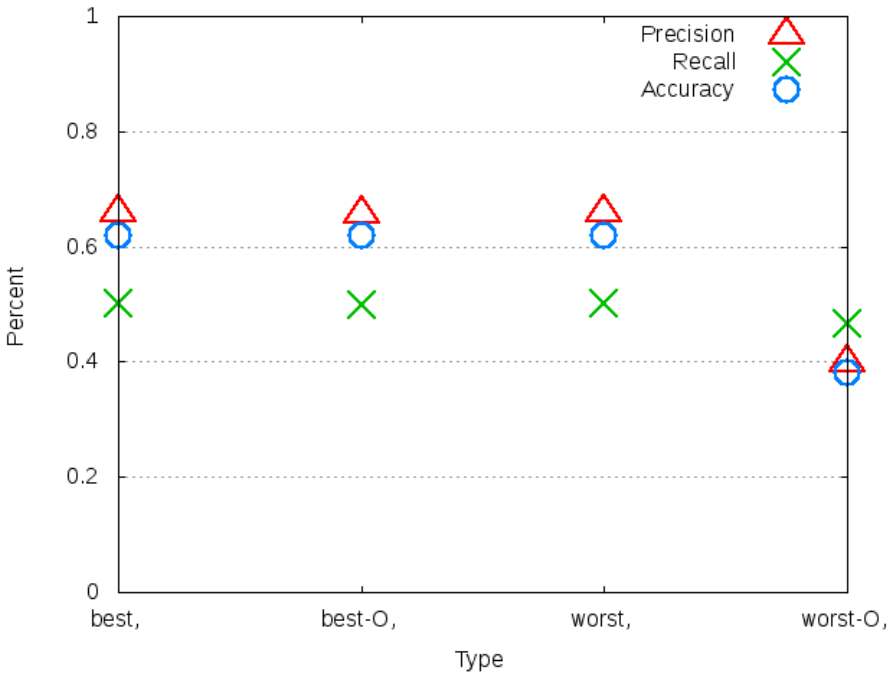


Figure A.137: Oversampling for tempto using SVM

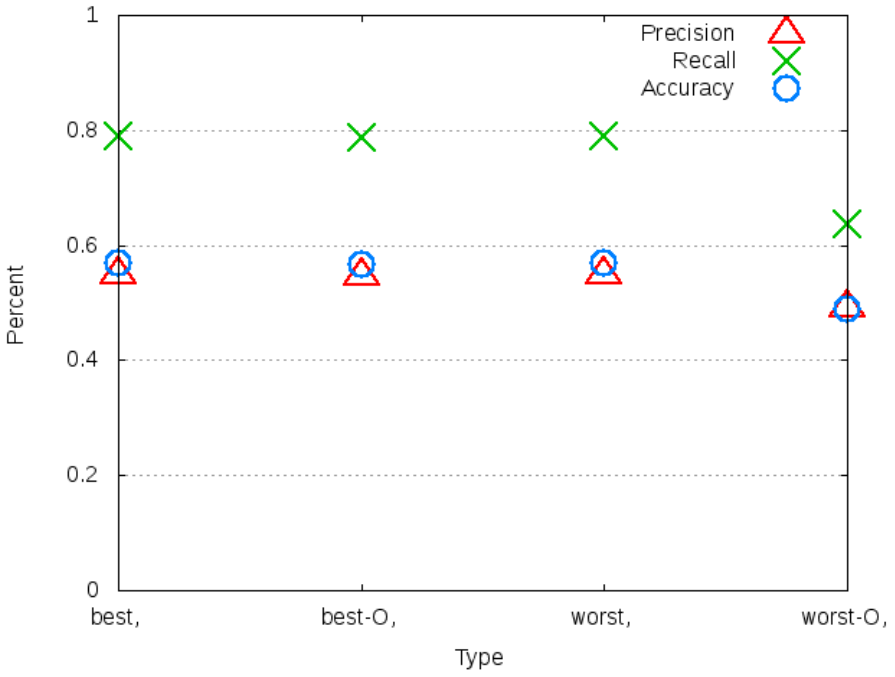


Figure A.138: Oversampling for yardstick using SVM

A.3.2 Random Forest

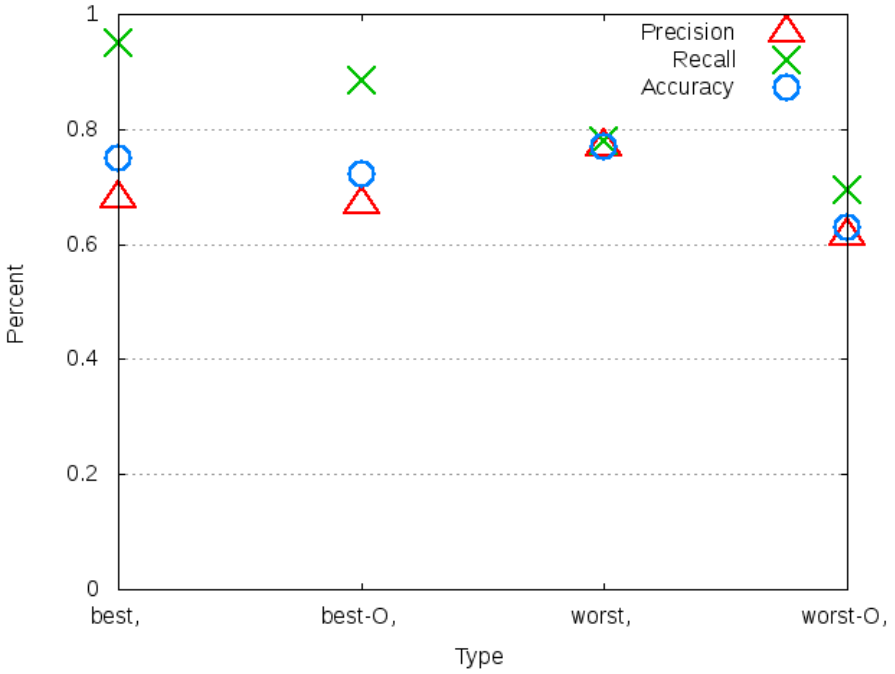


Figure A.139: Oversampling for acra using RF

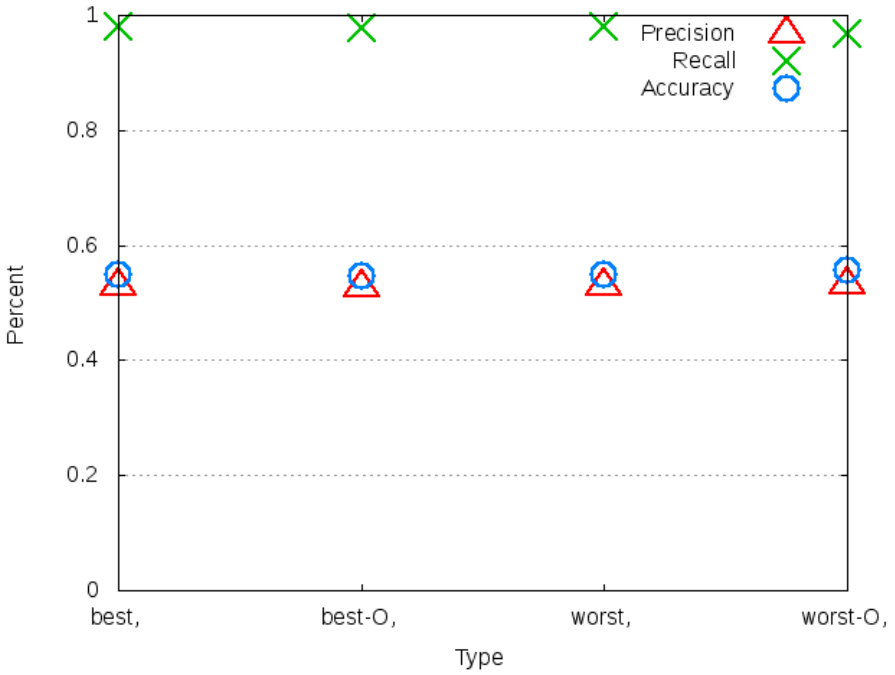


Figure A.140: Oversampling for arquillian-core using RF

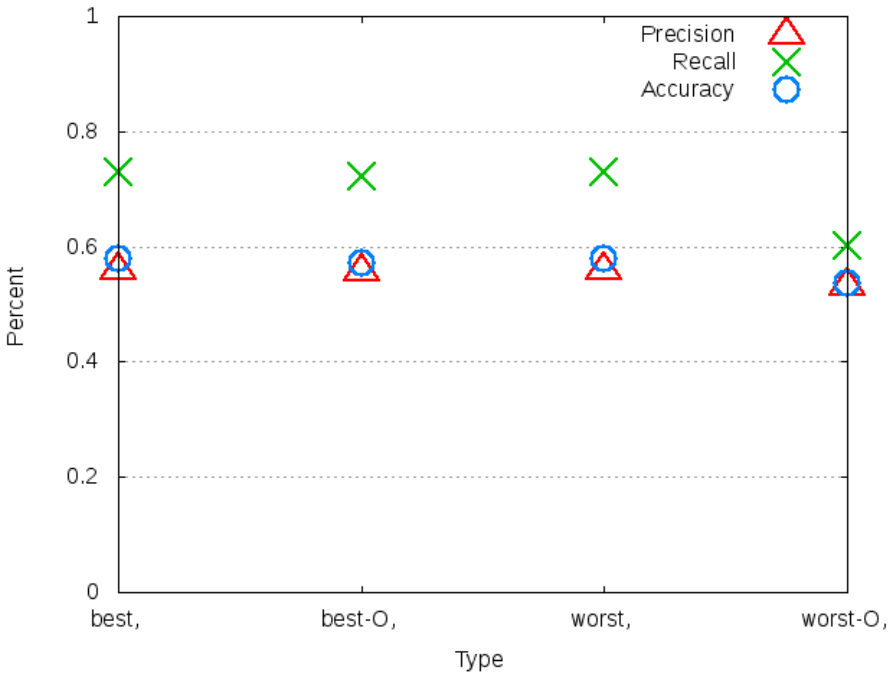


Figure A.141: Oversampling for blockly-android using RF

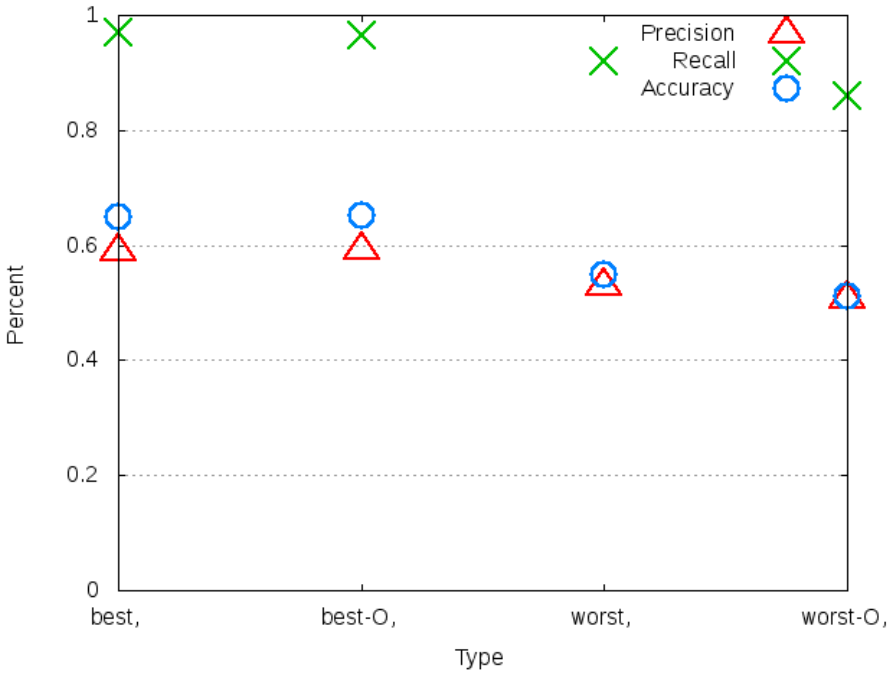


Figure A.142: Oversampling for brave using RF

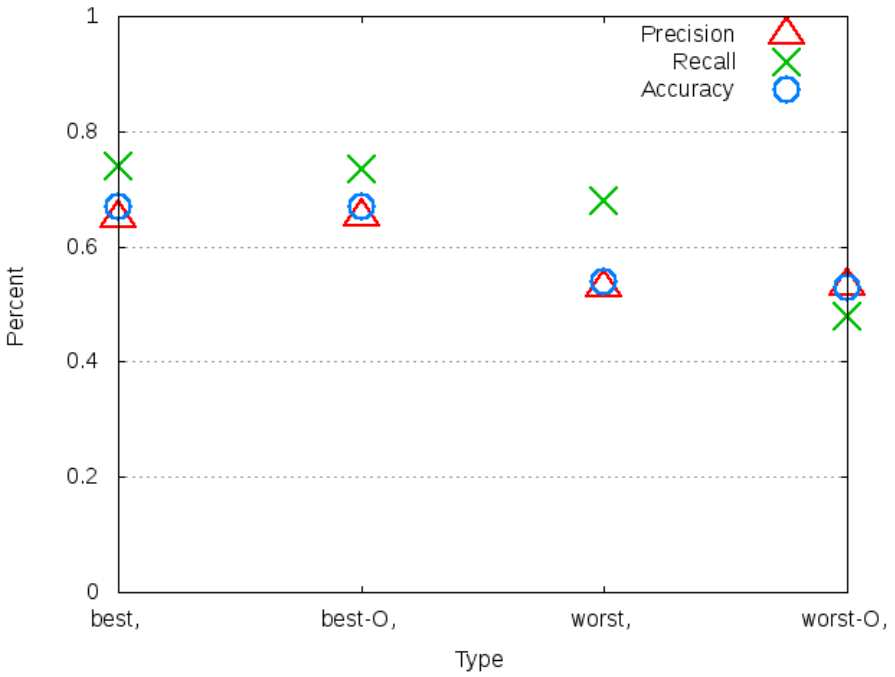


Figure A.143: Oversampling for cardslib using RF

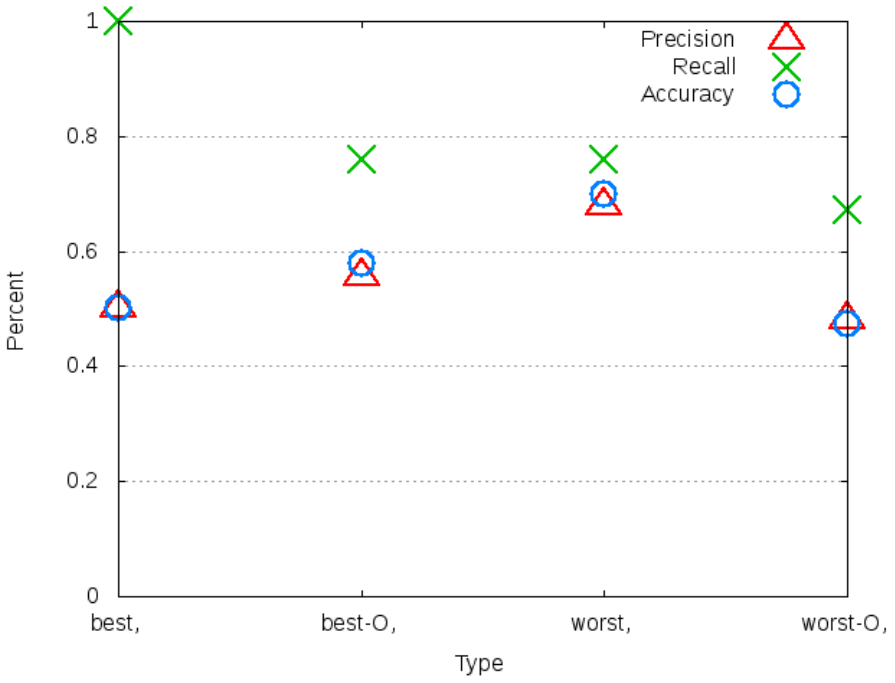


Figure A.144: Oversampling for dagger using RF

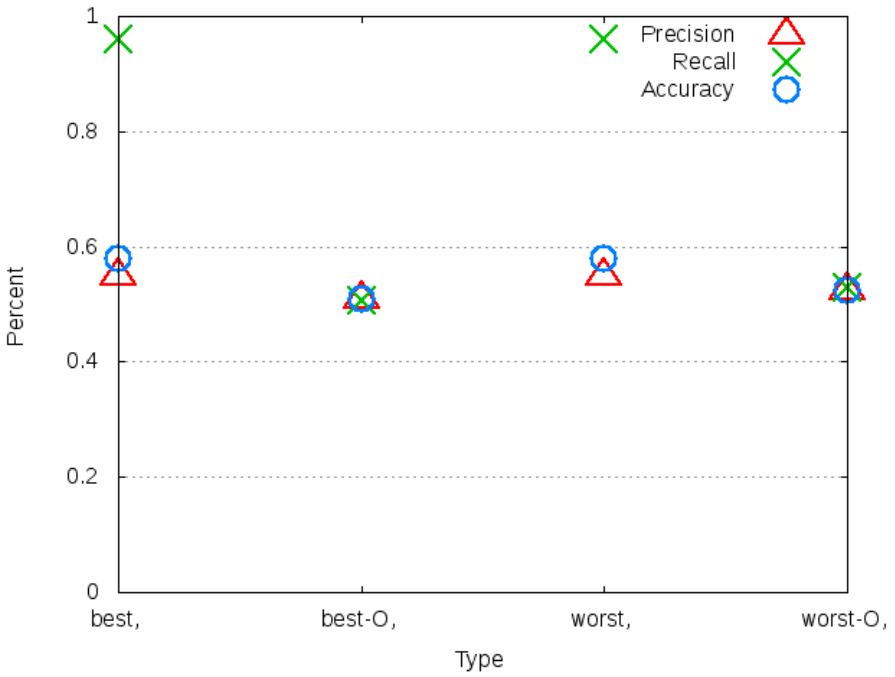


Figure A.145: Oversampling for deeplearning4j using RF

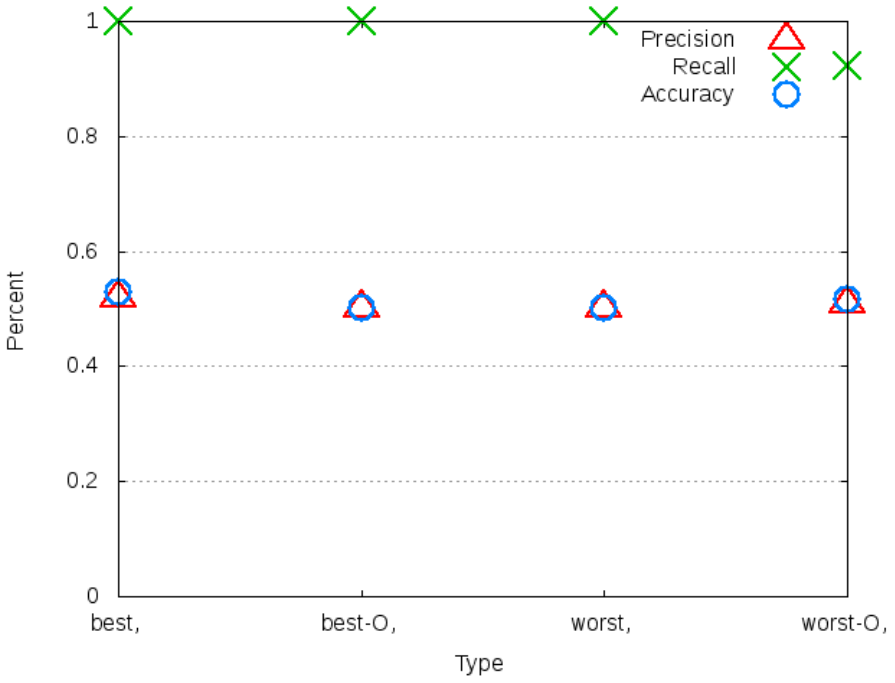


Figure A.146: Oversampling for fresco using RF

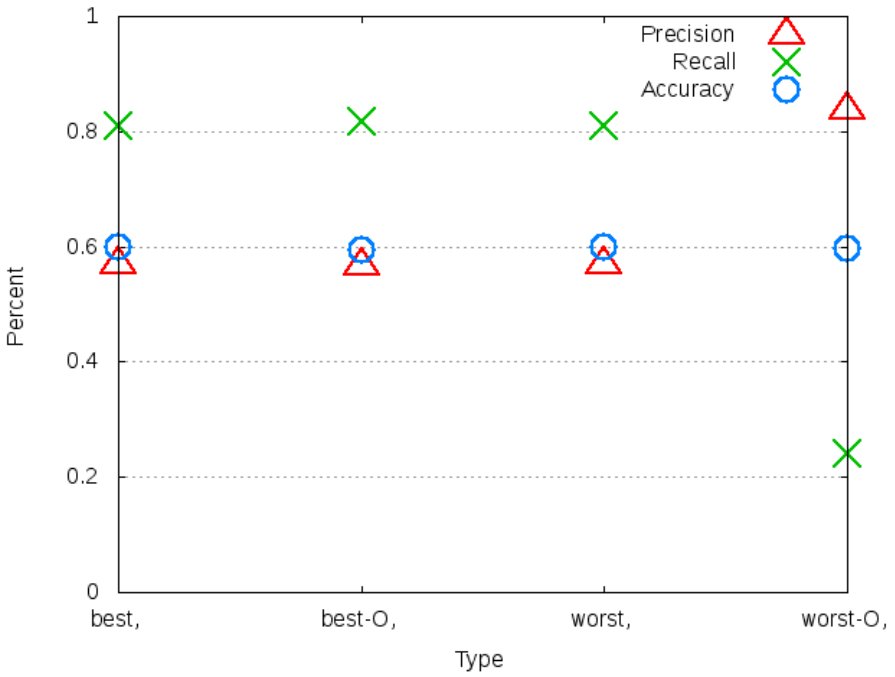


Figure A.147: Oversampling for governator using RF

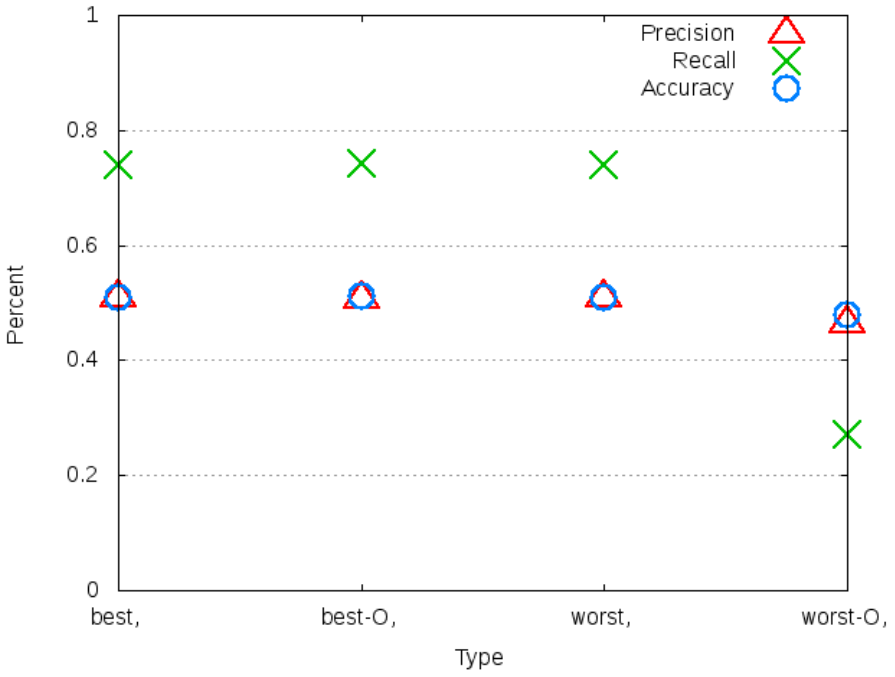


Figure A.148: Oversampling for greenDAO using RF

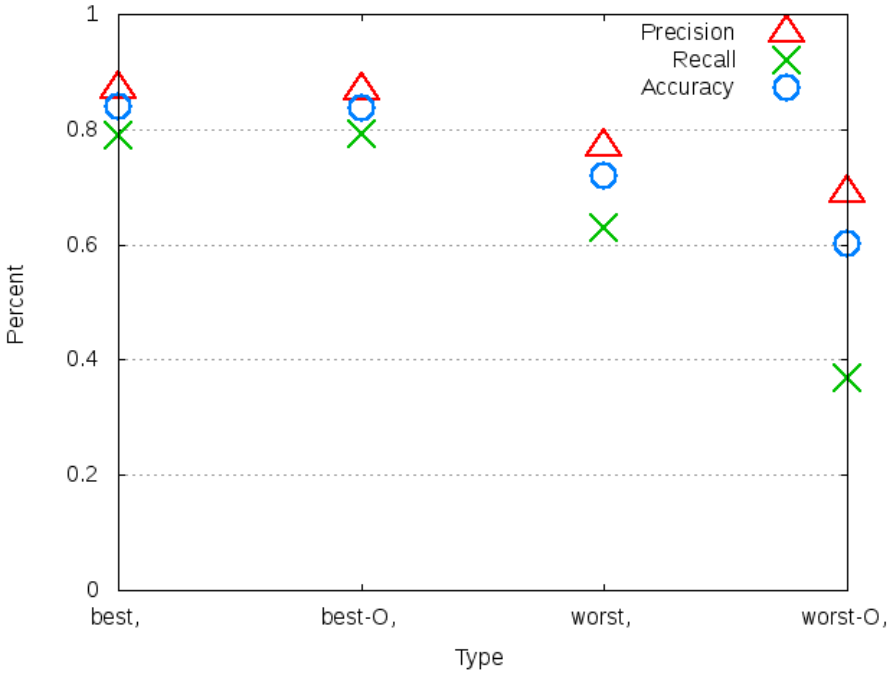


Figure A.149: Oversampling for http-request using RF

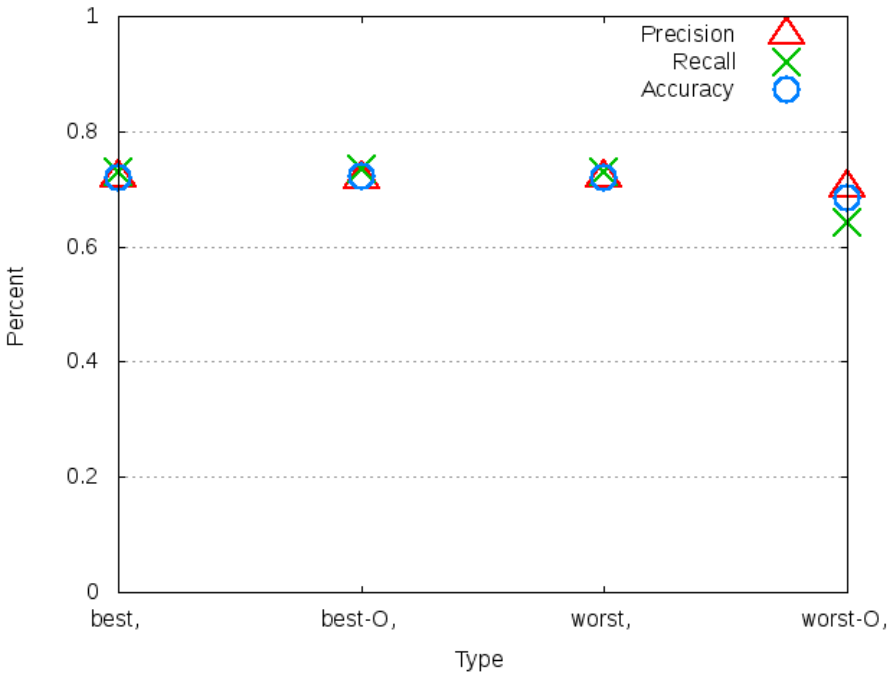


Figure A.150: Oversampling for ion using RF

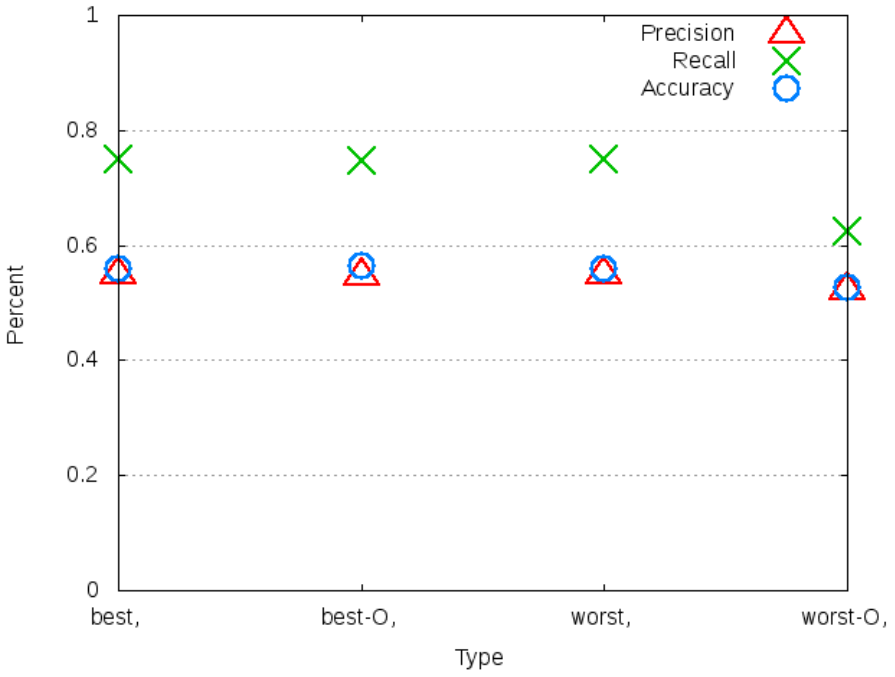


Figure A.151: Oversampling for jadx using RF

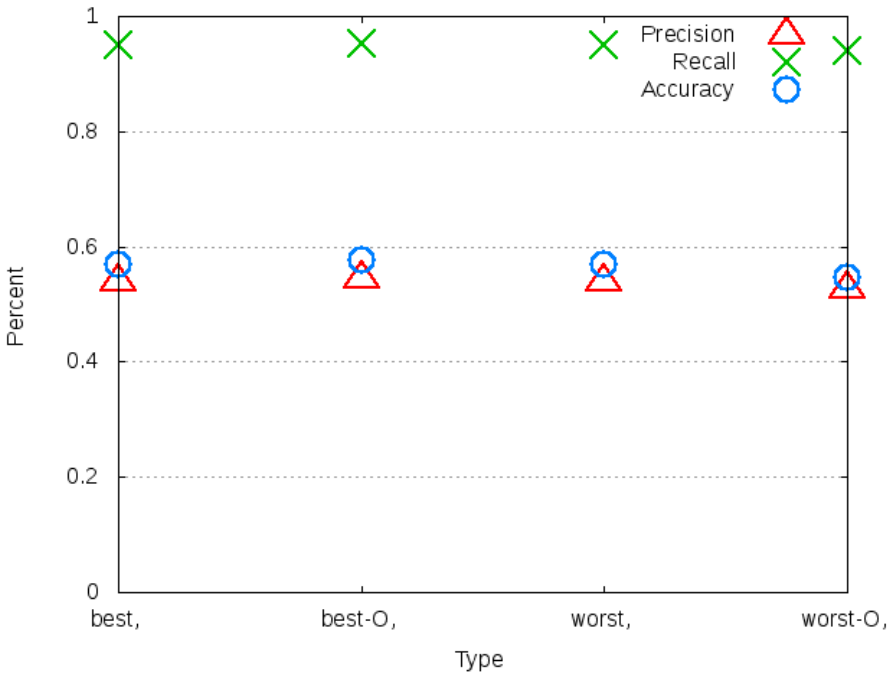


Figure A.152: Oversampling for mapstruct using RF

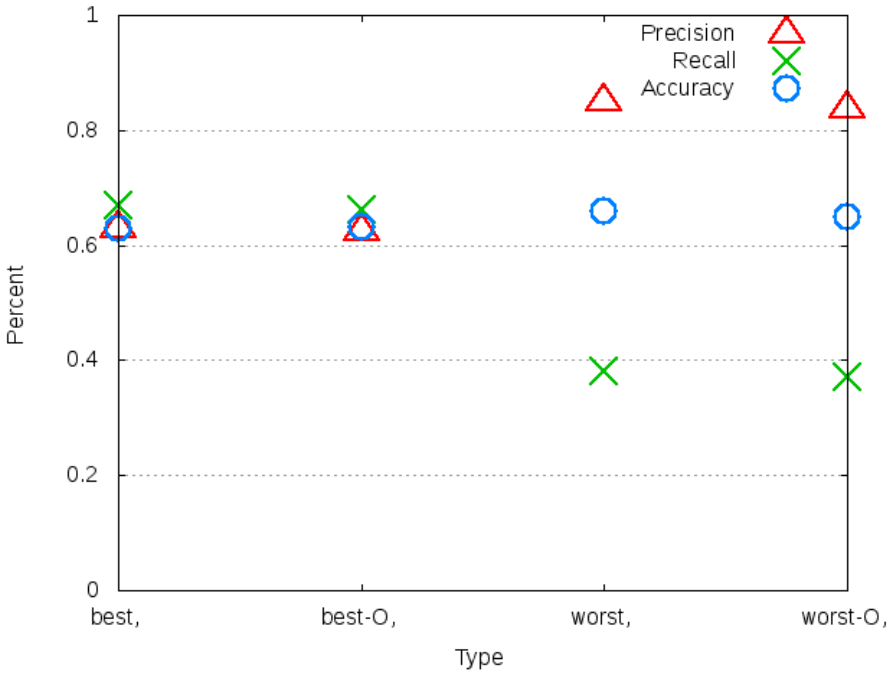


Figure A.153: Oversampling for nettosphere using RF

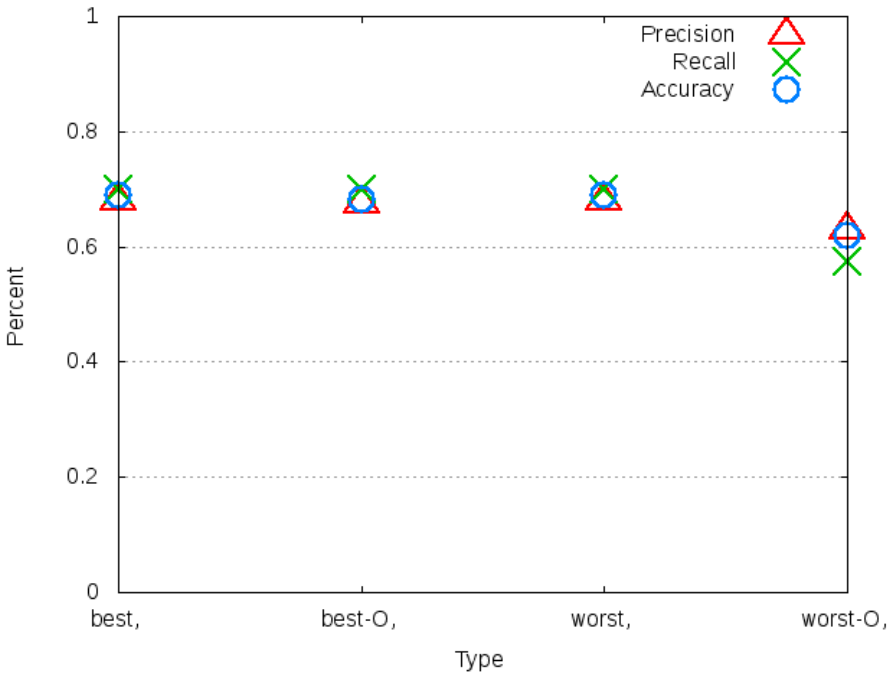


Figure A.154: Oversampling for parceler using RF

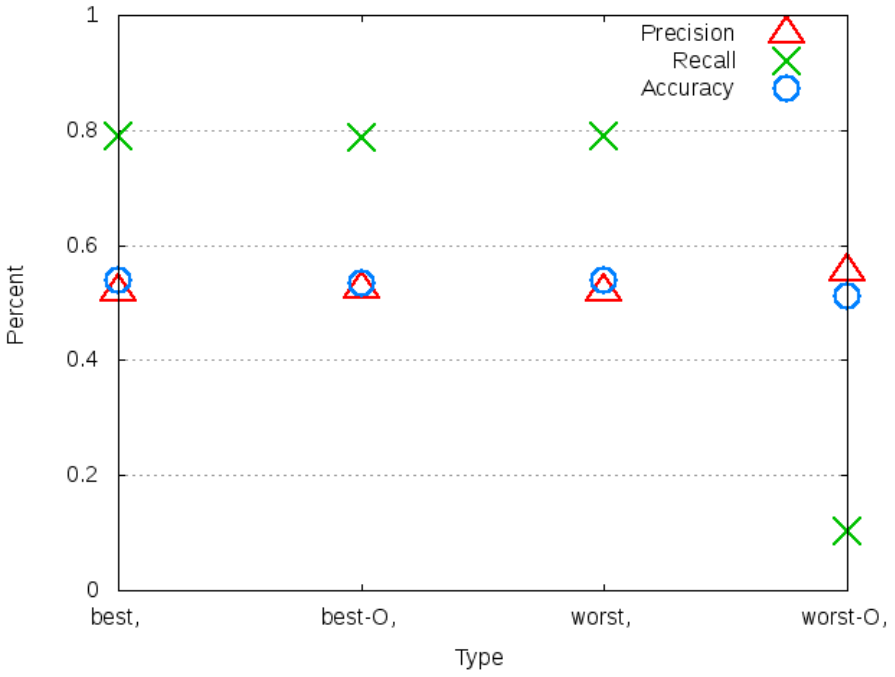


Figure A.155: Oversampling for retrolambda using RF

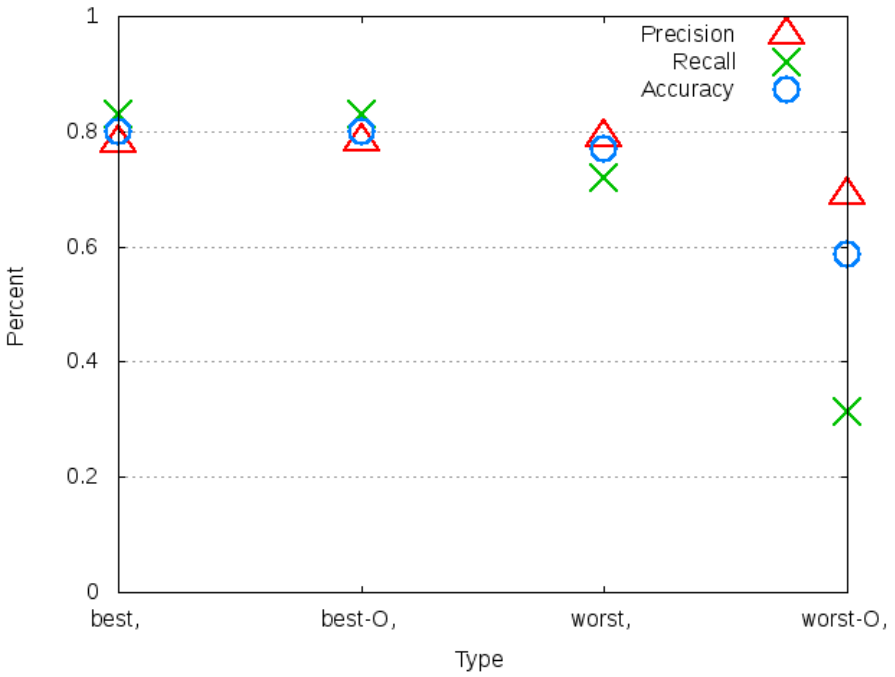


Figure A.156: Oversampling for ShowcaseView using RF

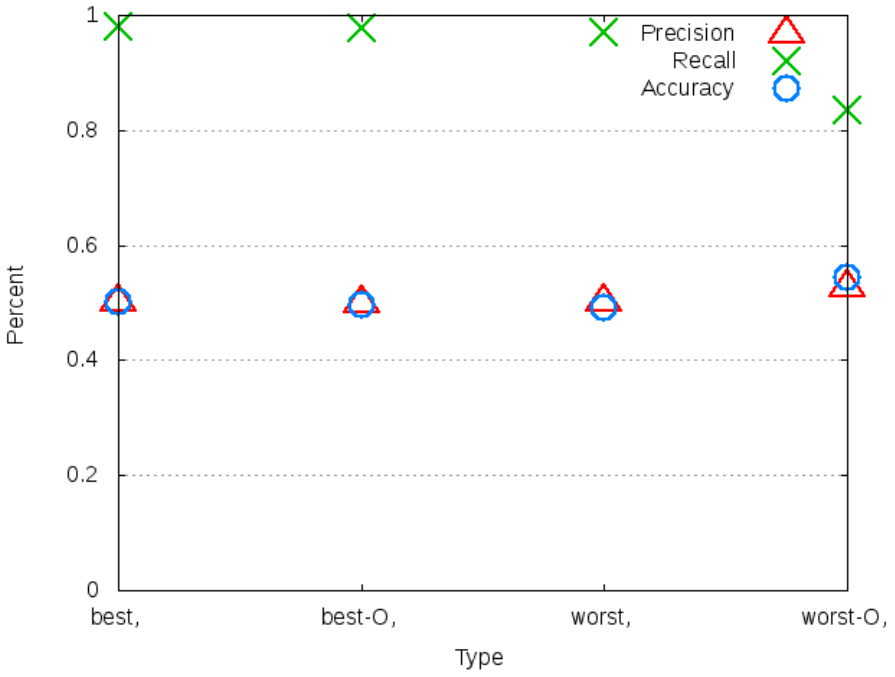


Figure A.157: Oversampling for smile using RF

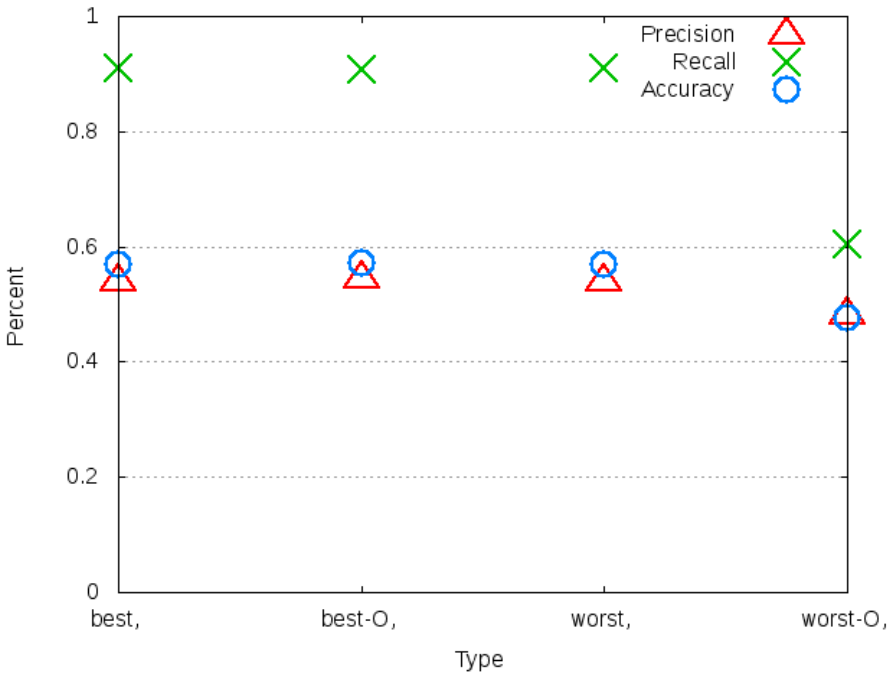


Figure A.158: Oversampling for spark using RF

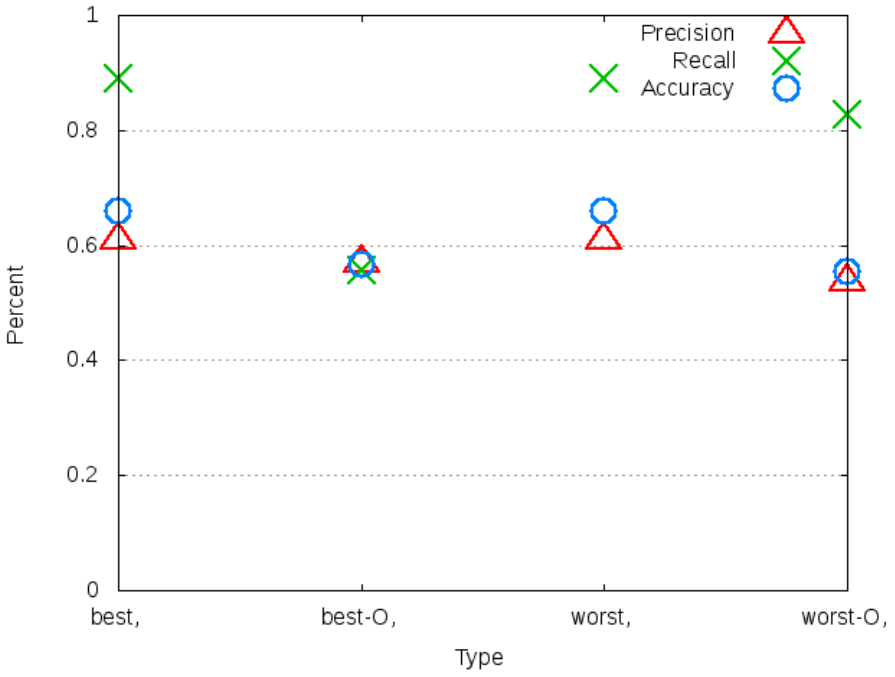


Figure A.159: Oversampling for storm using RF

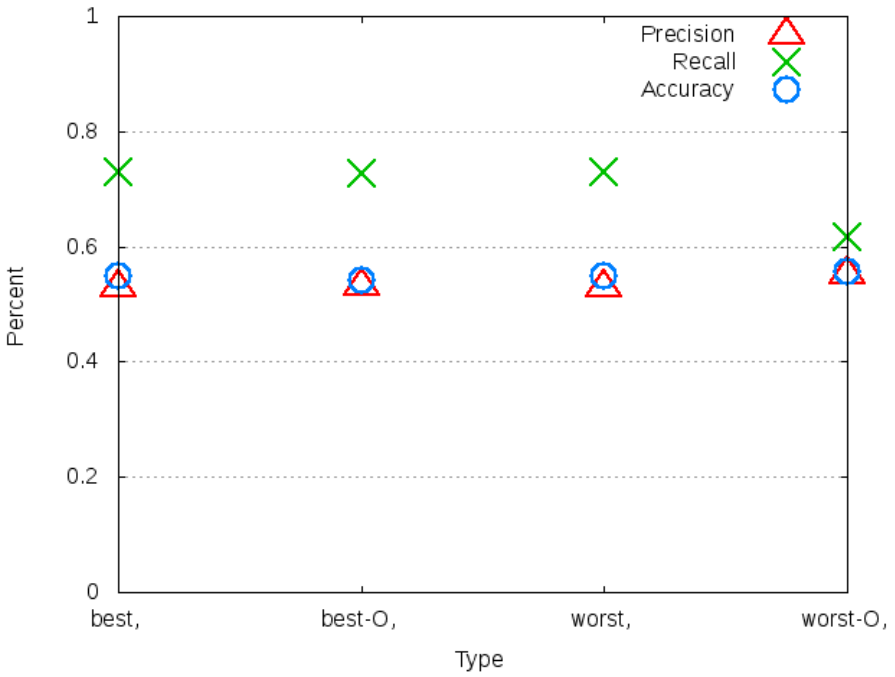


Figure A.160: Oversampling for tempto using RF

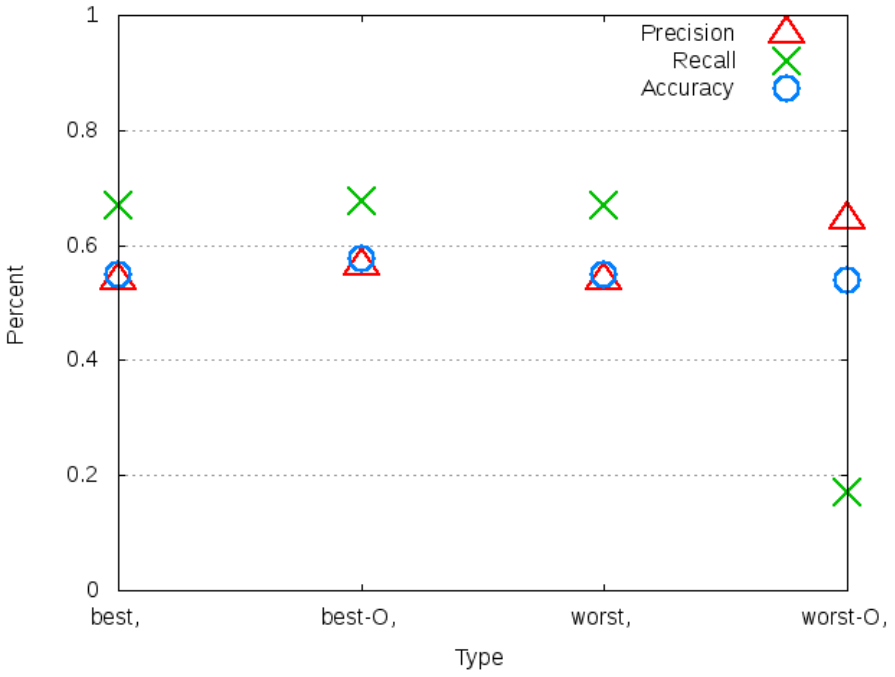


Figure A.161: Oversampling for yardstick using RF