Using Hoeffding sampling and project elimination to make Bellwether dicovery faster

Akshay Nalwaya
North Carolina State University
Raleigh, North Carolina
analway@ncsu.edu

Sanjana Kacholia
North Carolina State University
Raleigh, North Carolina
skachol@ncsu.edu

Shantanu Sharma
North Carolina State University
Raleigh, North Carolina
ssharm34@ncsu.edu

ABSTRACT

In software engineering, the project data is continuously updated and augmented. Prediction models build from these projects become increasingly varied as the number of projects increased and ultimately resulting in changing results. This problem of conclusion instability in software engineering, can be mitigated by using *"Bellwethers"*. It helps to build quality software prediction models. This problem was extensively researched in paper Bellwethers [1]. Bellwethers are used as baseline method for transfer learning and then this baseline is used for comparing future models.

So, in this paper we explore alternative methods to make the task of identification of bellwethers project in a group of projects faster for the defect prediction domain. An O(N^2) approach was presented in the Bellwethers paper and we try to explore the applicability of Hoeffdings bounds to sample the training set and experiment various combinations in the train and test sets. In addition to this, we also try to prune projects which are unlikely to be a candidate for bellwether project.

Keywords: Defect Prediction, Bellwether, Hoeffdings Bounds, Transfer Learning

1 INTRODUCTION

Bellwethers

The *bellwether effect* described in [1] states that when a community works on software, then there exists one exemplary project, called the bellwether, which can make predictions for the others.
The model built using bellwether project can serve as a baseline model for constructing different transfer learners in various domains of software engineering.

Importance of bellwether identification

If there is insufficient data, *transfer learning* can be used by data miners and use lessons learned from one project and apply them on another project.

Since the probability of having a defective code and a non-defective code is not similar, the SE data is often imbalanced and difficult to get. In such cases, Bellwethers method presents a simple solution - instead of exploring all available data, find one data set that may offer stable conclusion over longer period of time. Bellwethers [**?**] shows existence of such projects in SE datasets and strives to find them. It is true that bellwethers, with such simplicity, are always better than other complex algorithms used for similar applications. At the same time, it has been shown that bellwethers are capable of outperforming some of the more complicated algorithms.

Current Solution

Current methodology of bellwether identification is an N^2 algorithm that tends to evaluate each project against every other project. Though this is a simple process we believe that this approach needs investigation. We aim to find a method that identifies bellwether project in much lesser time without making the identification process much complex. Also, it should be scalable when number of projects is much larger that the ones explored in the current approach. We can also sample each dataset to reduce the training or testing dataset. This should lead to significant reduction in the dataset size.

Proposed Solution

In this paper we do various experiments with sampling the training and testing datasets. In addition to just sampling the datasets, we also compare this sampling method with idea of eliminating a project altogether without having to test it on all the projects. First and foremost we explored sampling methods available and zeroed on Hoeffdings bound method. Since we did not get significant improvement in the run times with the approach we then explored the ways to do project elimination.

Challenges faced

The goal of the project was to reduce the time taken to suggest bellwether, with that in mind we explored various algorithms and techniques to achieve results. The challenge we faced was either focus on reducing number of projects (N) or sample the datasets with various techniques. In the end we decided to explore other in isolation and see which approach among the two is better.

Overview of Results

The results of our experiments are documented in the results section. We see that project elimination is gives us the t est results with a scale of 6 with respect to the baseline. The sampling methods though gave good results in terms of dataset required to reach the baseline scores (we needed 8%-10% data) but gain in terms of time was negligible. This is the subject of further investigation.

2 RESEARCH QUESTIONS

Our study concentrated on finding answers to the following three research questions

RQ1: Can we predict which dataset is bellwether?
In many experiments that we perform we investigate if we can find the bellwether for given set of projects with sampling of both training and testing datasets.

RQ2: Can we reduce the time to find bellwether by reducing the size of data ?
Another problem worth investigating is that if we can reduce the

time to compute the bellwether using project elimination. We explore an algorithm that explores the possibility of not requiring to test current project against every other project. Since we need to find a winner project, it may just be possible.

RQ3: Does sampling data based on Hoeffding sampling outperforms idea of project elimination?
As mentioned above we try different experiments to check which method i.e. sampling or project elimination gives better result in terms of predicting Bellwethers and among them which can outperform other vis a vis better result stability or reduction in time for the experiment.

3 METHODOLOGY

In this section we describe various steps done and try to address why those were required. We cover dataset description, generation of baseline and evaluation criteria.

3.1 Data Set Description

Since we limited out work to finding Bellwethers for defect measures we relied on dataset gathered by Jureczko. The data set contains defect measures from several Apache projects. The dataset comprises of data from 10 different projects. This dataset contains records the number of known defects for each class using a post-release bug tracking system. The classes are described in terms of 20 metrics. Each dataset in the Apache community has several versions. We merged dataset across different version to create bigger dataset.

Following are the details of the dataset for each project. All projects had 20 features. Since the class variable was continuous we did some pre processing to convert to binary. This was done because our objective was to find whether for a given instance, will it have a bug or not and not the number of bugs identified. So we mapped all the instances having at least 1 bug as positive class while those not having any bug as negative class.

S No	Project	Rows
1	ant	1692
2	camel	2784
3	ivy	704
4	jedit	1749
5	log4j	449
6	lucene	782
7	poi	1378
8	velocity	689
9	xalan	3320
10	xerces	1643

3.2 Baseline

There are many binary classifiers to predict defects, the Bellwethers [1] cites studies on defect prediction and follows the use of Random Forests for defect prediction over several other methods. For sake of simplicity and effective comparison (if required) we decided to use the Random Forest Classifier.
Baseline calculation is a straight forward task - for each project in the community *train* the model, and *test* it against every other project and compute G-Scores for each of the test iteration. The project with the best median value of the G-Score is declared the Bellwether.

Algorithm 1 Baseline

```
start time = current time
for project in projects
    read data
    train random classifier
    for testproj in projects
        if testproj not equal to project
        make predictions on testproj
        calculate gscore
        append g to a table
end time = current time
runtime = end time - start time
```

3.3 Evaluation

The dataset under consideration has binary class labels, with the records belonging to either positive or negative class. The instances of projects having defects (one or more) are assigned positive class while those without defects are assigned negative class implying no defect was found in that instance.

There are various metrics that can be derived from the confusion matrix obtained as a result of testing the classifier on each of the projects. Different measures of model evaluation are summarized below:

Standard Measures of Evaluation

Accuracy: It is the percentage of instances of the dataset that have been classified correctly by the model. It emphasizes on correct classification of both positive and negative classes equally. The mathematical formula for accuracy is,

$$accuracy = \frac{true\,positive + true\,negative}{total number of instances}$$

Precision: It talks about how precise your model is, meaning it shows what fraction of instances that are predicted positive, are actually positive. So a model with low precision would imply that either a there was a large number of false positives in the model or the number of true positives was very low.

$$precision = \frac{true\,positive}{true\,positive + false\,positive}$$

Recall: It calculates how many of the actual positive instances have been correctly captured by the model (true positives). It is also denoted by *pd* or the *probability of detection.*

$$recall = \frac{true\,positive}{true\,positive + false\,negative}$$

False Alarm: As the name suggests, this metric gives the percentage of negative instances that were erroneously predicted as positive instances. It is also denoted by *pf.*

$$pf = \frac{false\,positive}{false\,positive + true\,negative}$$

Each of the metrics we discussed above are used for model evaluation depending on the application and the type of data. For instance, if one aims to increase the recall for a model, then it might also increase the false alarm (pf) of the model. Similarly, there is a kind of inverse relationship present in between precision and recall. If one tries to increase the precision of a model, then the recall might have to be compromised with.

Class Imbalance in classification problems is a scenario where classes are not represented equally. Most classification datasets do not have an equal representation of the classes and often such class imbalance needs a careful handling. Slight variations in the class distributions can be ignored but a significant variation needs the be taken into account. There are several ways of handling the class imbalance and most common among them are

- Collect more data
- Change performance metric
- Re-sampling dataset

Why Not Accuracy

In the cases discussed above, accuracy can often be misleading. At times it may be desirable to select a model with a lower accuracy because better predictive power on the problem. For example, in a problem where there is a large class imbalance, a model can just predict the value of the majority class for all predictions and achieve a high classification accuracy, the problem is that this model is not useful in the problem domain.

G-Score

We use the **G-Score** as a metric for evaluating performance of classifier in this case of class imbalance. It combines recall (pd) and false alarm rate (pf). The Bellwethers [1] cites studies which suggest that such a measure is justifiably better than other measures when the dataset has imbalanced distribution in terms of classes. Hence we are using G-Score in this paper as well.
G-Score is measured as follows:

$$G = \frac{2 \times Pd \times (1 - Pf)}{(1 + Pd - Pf)}$$

It is clear from this formula that models having higher G-Scores are better.

4 EXPERIMENTS

4.1 Hoeffdings Bounds

Let us say that we have N points with which to test a given model. If we were to test a model on all of them, then we would have an average error that we will call E_{true}. However, if we only tested the model on ten points, then we only have an estimate of the true average error. We call the average after only n points (n<N) E_{est} since it is an estimate of E_{true}. The more points we test on (the bigger n gets), the closer our estimate gets to the true error. How close is E_{est} to E_{true} after n points? Hoeffdings bound lets us answer that question when the n points are picked with an identical independent distribution from the set of N original test points. In this case, we can say that the probability of E_{est} being more thanaway from E_{true}

$$Pr(|E_{true} - E_{est}| > \epsilon) < 2e^{-2n\epsilon^2/B^2}$$

where B bounds the greatest possible error that a model can make [2]. This bound does not make any assumptions other than the independence of the samples

4.2 Sampling

S No	Sampled Training Data	Different % of Training Data for each Testing Data
1	No	No
2	Yes	Yes
3	Yes	No

4.2.1 ***Approach 1.*** In this approach, our motive was to decrease the time required to find Bellwethers by decreasing the training dataset. Each dataset was taken, and the samples were added to the training data set. We start with 5% of the data set and at each iteration, 1% of the data is added. Model is trained on this data at each iteration and tested on the other projects. The point at which we are 95% confident that our estimate of the running g score is within the epsilon of baseline g score, is noted. The loop breaks for that test project and runs for all the remaining test projects. Once, we hit hoeffding bound for each test project, a similar exercise is run for the remaining projects. This helped us reduce the training data considerably. However, the fraction of data being sampled for each test project was different. Hence, the time taken to run the process was not reduced. The fact that sampling the training data set for each test set was increasing the time of execution led us to devise another approach.

Algorithm 2 Approach 1 Algorithm

```
start time = current time
for project in projects
    read data
    for testproj in projects
        if testproj not equal to project
            sample data upto hoeffding bound
            train random classifier
            make predictions on testproj
            calculate gscore
            append g to a table
    end time = current time
runtime = end time - start time
```

4.2.2 ***Approach 2.*** In order to eliminate the time taken for sampling the training data differently for each dataset, we took the maximum of the percentage of data required by any test project. This did not just decrease the training dataset but impacted the time taken to find bellwethers. The time taken to find Bellwethers was reduced by 4 times. In this approach, the sampling of data was done just once, and with the reduced training data set, predictions are made. G score calculated is almost similar to the values calculated through baseline method.

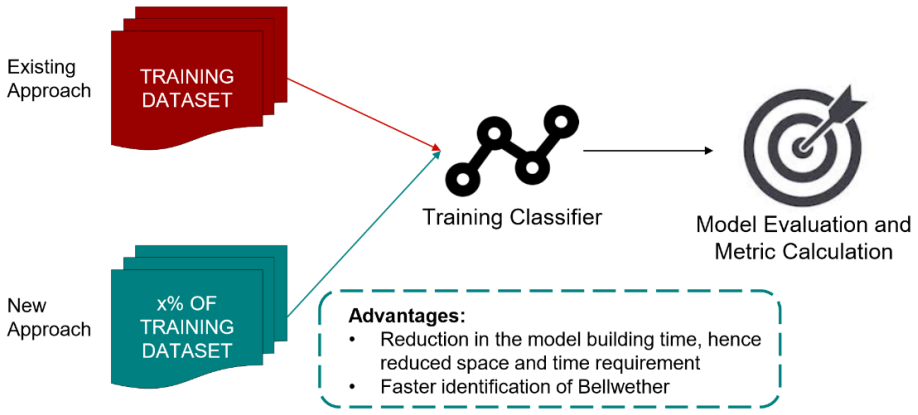

Figure 1: High level process flow of current process and comparing it with our Hoeffding Bounds process

Algorithm 3 Approach 2 Algorithm

```
start time = current time
for project in projects
    read data
    frac = get max % of training data
    sample data for frac
    train random classifier
    for testproj in projects
        if testproj not equal to project
            make predictions on testproj
            calculate gscore
            append g to a table
    end time = current time
runtime = end time - start time
```

4.3 Project Elimination

Algorithm 4 Project Elimination Algorithm

```
for each project do
    load X_train, y_train
    train classifier
    set threshold g-score
    for all other projects
        load X_test,y_test
        predict
        compute g-score
        If eliminate-count > 2:
            break
    If g-score < threshold and
            num projects tested >=3:
            g-score=0
    append results, g-score
return results
```

In the above experiments, we focused on using Hoeffding Bounds for efficiently sampling the datasets for all the projects. Depending on the approach, we experimented with sampling only the training data, or testing data, and then sampling both: training and testing data. The end goal was to try and reduce the number of records to be used for training and testing purposes. This doesn't reduce the number of projects but it reduces the constant term in run-time complexity analysis, i.e., in cN^2 the term c is reduced.

In this experiment, our objective was to explore if we can eliminate some projects and hence try to reduce the value of N in the complexity expression. We studied the baseline results for all the projects and found that some projects which consistently had low values for G-score for prediction on other projects continued this trend for all the projects. This accounted for a large amount of time being spent on training and testing of projects that were not the candidates for bellwether. So, such projects should be eliminated without spending time in using such projects for testing for all other projects.

Based on the baseline G-score values we decided a threshold value for the G-score. This threshold value should be satisfied for all the projects in order to be considered as a candidate bellwether project. The central value of G-score distribution for all the projects was chosen as the threshold value. Mean, median and mode are the most commonly used measures of central tendency. We chose median as the representative value of the central tendency for bellwethers since mean is prone to be influenced by the presence of outliers in the data (G-score of projects in this case). One project with very low G-score could bring down the threshold G-score and lead to increased processing time. On the other hand, median is not affected by some outliers in data, and hence is a better measure for this case.

Once this threshold was decided, we started with training a Random Forest Classifier on each of the projects. This trained random forest classifier is used for testing on all the other projects in a sequential manner. The G-score values for each iteration is recorded for each of the trained classifier model. The current project is eliminated if the following two conditions are satisfied:

Trained on	Tested on									
	ant	camel	ivy	jedit	log4j	lucene	poi	velocity	xalan	xerces
ant	-	5	6	7	8	9	10	11	12	13
camel	5	-	6	7	8	9	10	11	12	13
ivy	5	6	-	7	8	9	10	11.5	12	12.5
jedit	5	6	7	-	8	9	10	8	12.5	12
log4j	5	6	7	8	-	9	10	11	12	13
lucene	5	6	7	8	9	-	10	11	12	13
poi	5	6	7	8	9	10	-	11	12	13
velocity	5	6	7	9	7	10	11	-	12	13
xalan	5	6	7	8	8.5	10	11	12	-	13
xerces	5	7.5	7	8	8	9.5	11	12	13	-

Figure 2: Used Hoeffding bounds to find the percentage of data required for training on a single project using random sampling and testing on all the other projects.

- ***Condition 1:*** Project is tested on at least $1/3^{rd}$ of the projects
- ***Condition 2:*** Mean of G-score value is less than the specified threshold value

All the projects satisfying these conditions are tested for the other projects until they violate these conditions or it has been tested on all available projects. The projects violating these conditions are pruned and also removed from the list of candidate bellwether projects. This serves as the early stopping rule to avoid testing on all projects and efficiently reducing the number of projects used for testing.

The G-score values are then aggregated for projects that have not been eliminated. Median value of G-score is taken and reported as the G-score for each project and the project with highest mean value of G-score for testing on other projects is termed as the bellwether project for the given set of projects.

5 RESULTS

RQ1: Can we predict which dataset is bellwether ? Predicting bellwethers using baseline method, generates the following G-score. The G-score found by various approaches are similar and does not vary much. The bellwether dataset can be predicted using the G-score and poi, proves to be the data set with the best G-score. However, 'xalan' and 'lucene' have median G-score comparable to that of 'poi'. We have tried three different approaches to predict Bellwethers and all the three approaches, 'poi' has the highest G-score throughout.

RQ2: Can we reduce the time to find bellwether by reducing the size of data ? For research question 2, we found the point when hoeffding bound is hit. Instead of the whole training data set, only a percentage of data set sample can be taken. The results are shown in the table below. All results were calculated after running for 30 iterations. The data for training can be reduced a lot but sampling the data again and again consumes a lot of time. Hence, we took the maximum amount of training data required to hit the hoeffding bound for any test project and sampled training data acoordingly. The sampling of data for each training set is just done once, considerably reducing run time. The results show that the run time has reduced more than 4 times. Tables below show the percentage of

Project	Baseline	Hoeffding Bound	Sampled Training Data
ant	0.18	0.18	0.19
camel	0.24	0.25	0.24
ivy	0.09	0.12	0.12
jedit	0.04	0.03	0.04
log4j	0.34	0.34	0.32
lucene	0.52	0.52	0.51
poi	0.61	0.62	0.61
velocity	0.49	0.49	0.49
xalan	0.56	0.58	0.57
xerces	0.42	0.43	0.43

Figure 3: Median G Scores for Bellwether Prediction

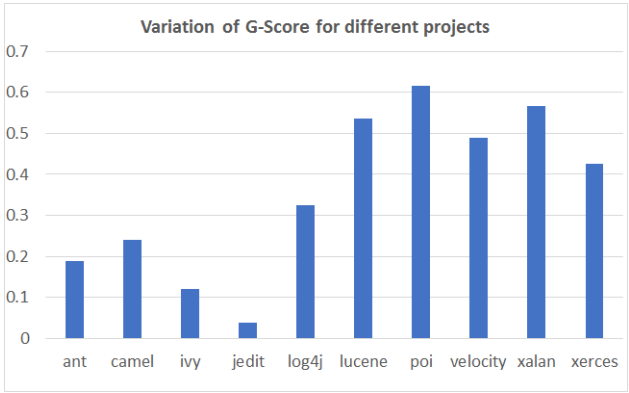

Figure 4: The median g-score for 'poi' was highest amongst all the projects used for defect prediction.

data used for training, their updated G-score and the runtimes for each approach.

RQ3: Does sampling data based on Hoeffding bounds outperforms idea of project elimination?

Based on the values of average runtimes in the figure 5, it is evident that eliminating projects definitely reduces the time required for finding bellwether project when compared with the baseline runtime obtained by round robin training of all projects against

Projects	% data for Training
ant	13
camel	13
ivy	12.5
jedit	12.5
log4j	13
lucene	13
poi	13
velocity	13
xalan	13
xerces	13

Figure 5: Approach 2 - Percentage of data required for Training for Bellwether Prediction

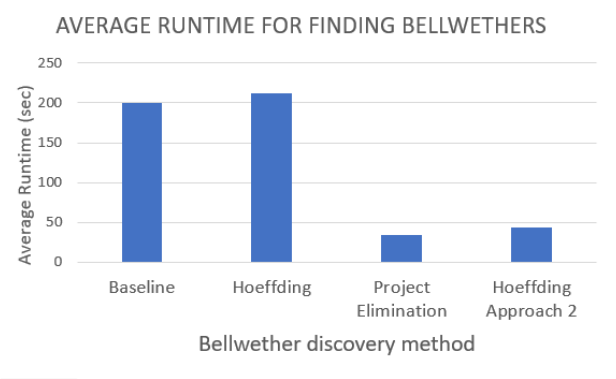

Figure 6: Average Runtime values for different approaches of finding Bellwether project

S No	Experiment Name	Time Taken(secs)
1	Baseline	199.32
2	Hoeffding 1	211.57
3	Hoeffding 2	44.16
4	Project Elimination	34.63

Table 1: Comparing run-times of all the approaches for finding Bellwethers.

Projects Eliminated				
ant	camel	ivy	jedit	log4j

every other project. In addition to beating the runtime of baseline method, project elimination proved to be even better than sampling data using Hoeffding bounds.

6 FUTURE WORK

Based on the results achieved in this project work, there are some more algorithms, and methods which can be explored to get even much better results. Some of the questions for which more exploration can be done are discussed below.

Projects	G score
lucene	0.54
poi	0.61
velocity	0.49
xalan	0.57
xerces	0.07

Figure 7: G-scores using Project elimination to predict Bellwethers

- *Explore alternative sampling methods*
  Can we implement more sampling methods instead of just Hoeffdings bounds and compare them? There is another sampling method named Bayesian Races which assume that data is normally distributed. This approach might also prove to be useful for reducing the time complexity associated with bellwether discovery.
- *Exploring project elimination with sampling*
  In this project we focused on evaluating sampling algorithms and eliminating projects in isolation. It would be interesting to see if we are able to achieve even better results by combining the idea of project elimination and data sampling.
- *Exploring with different repositories*
  For the scope of this project we limited ourselves to the use of Jureczko repository, but we can explore more dataset repositories and use them to explore our research questions. It would help us to see how well does our work apply to other repositories.
- *Adding Parallelism to code*
  Another important aspect of computation is parallelism built in languages and platforms. Our current work does not explore this but adding parallelism to the code should lead to lower latency.

7 CONCLUSION

In this paper, we have performed a thorough study of Bellwether discovery process and their importance in various software engineering domains. Our results show that we can make the process of identification of bellwether project in a repository much faster than the current round robin $O(N^2)$ approach.

We have showed that sampling the training and/or testing datasets helps in reducing the amount of time spent for bellwether discovery process. Data sampling not only helps to reduce the runtime of the code but also shows that efficiently selecting data from a project can help minimize redundancies and ensure that the classifier is trained only on the optimum percentage of training data. This comes at almost negligible loss of the model performance measured using G-score.

We also proved that sampling is not always the best method for reducing the time for bellwether discovery. We can prune projects to avoid the time spent on training projects which are highly unlikely to be prospective bellwethers. This helped to focus only on the projects that are capable of being the representative project for the

whole repository which essentially reduced the runtime by a some fraction of N, with N being the number of projects in the repository.

REFERENCES

[1] Tim Menzies and Rahul Krishna *Bellwethers: A Baseline Method For Transfer Learning* . IEEE Transactions on Software Engineering, April 2018

[2] Oden Maron and Andrew W. Moore *The Racing Algorithm: Model Selection for Lazy Learners.* Artificial Intelligence Review, February 1997, Volume 11