**RESEARCH STUDY**

1. Software Vulnerability Analysis and Discovery Using Machine-Learning & Data-Mining Techniques: A Survey

By: -

SEYED MOHAMMAD GHAFFARIAN and HAMID REZA SHAHRIARI,

Amirkabir University of Technology

ACM Computing Surveys Journal.

Vol. 50, No. 4, Article 56.

Publication date: August 2017.

**OBJECTIVE:** Software security vulnerabilities are one of the critical issues in the realm of computer security. Due to their potential high severity impacts, many different approaches have been proposed in the past decades to mitigate the damages of software vulnerabilities. Machine-learning and data-mining techniques are also among the many approaches to address this issue. In this article, an extensive review of the many different works in the field of software vulnerability analysis and discovery that utilize machine-learning and datamining techniques.

**APPROACHES USED BY VULNERABILITY MODEL**



Fig. 1. Two-fold categorization scheme of reviewed work

**OBSERVATIONS**

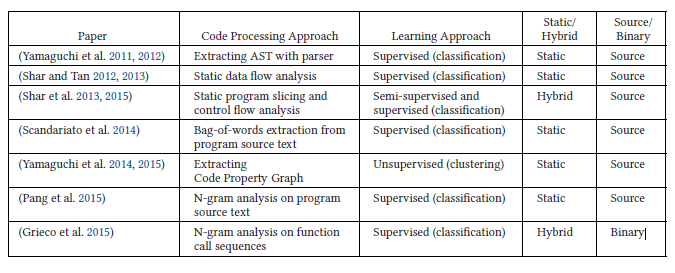
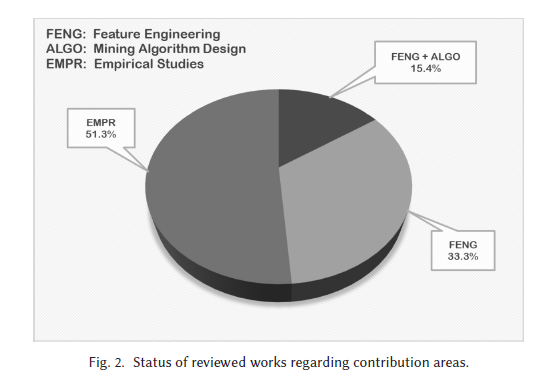


Table.1. Summary of Work on Vulnerable Code Pattern Recognition

**CONCLUSION AND FUTURE WORKS**

Data-mining and machine-learning techniques have been successfully used in many different application domains, including the domain of computer security. In this paper, we extensively reviewed previous work that applied data-mining and machine-learning techniques for software vulnerability analysis and discovery.



The main conclusion that we draw from the study of previous approaches is the yet immature state of research in the field of software vulnerability analysis and discovery using machine learning and data-mining techniques. Most of studies provide extended empirical case studies, which try to apply well-known techniques and best-practices from the fields of machine learning and datamining on software vulnerability data, to investigate the answer of some research questions.

1. ESTIMATING THE PREDICTION PERFORMANCE OF SPATIAL MODELS VIA SPATIAL K-FOLD CROSS VALIDATION

By: -

Jonne Pohjankukka, Tapio Pahikkala, Paavo Nevalainen & Jukka

Heikkonen

Department of Future Technologies, University of Turk

Publication date: July 2017

International Journal of Geographical Information Science

Vol. 00, No. 00

**OBJECTIVE:** In machine learning one often assumes the data are independent when evaluating model performance. However, this rarely holds in practise. Geographic information data sets are an example where the data points have stronger dependencies among each other the closer they are geographically. This phenomenon known as spatial autocorrelation (SAC) causes the standard cross validation (CV) methods to produce optimistically biased prediction performance

estimates for spatial models, which can result in increased costs and accidents in practical applications. To overcome this problem, a modified version of the CV method called spatial k-fold cross validation (SKCV), which provides a useful estimate for model prediction performance without optimistic bias due to SAC. We test SKCV with three real world cases involving open natural data showing that the estimates produced by the ordinary CV are up to 40% more optimistic than those of SKCV. Both regression and classification cases are considered in our experiments. In addition, we will show how the SKCV method can be applied as a criterion for selecting data sampling density for new research area.

**ATTEMPTS TO THE SOLUTION**

More specifically, SKCV attempts to answer the following two questions:

(1) What is the prediction performance of a model at a certain geographical

location when the closest data measurements used to train it lie at a given

geographical distance?

(2) Conversely, if the prediction performance is required to be at least a given

level, how dense data sampling grid should be used in the experiment area

to achieve it?

**AN ILLUSTRATION OF THE CONSIDERED EXAMPLE USED**

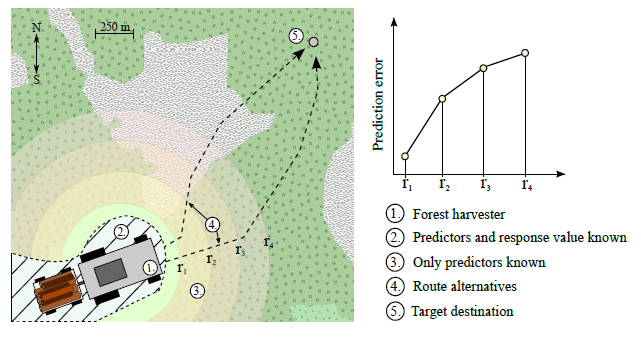
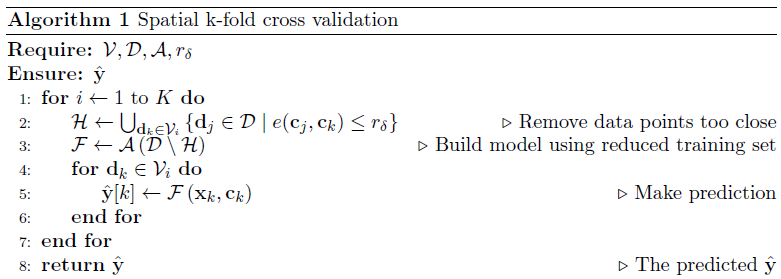


Figure .1. The forest harvesting example. The harvester driver needs to select an optimal

route to target destination. Due to SAC it is to be expected that the prediction error increases the further away we make point predictions. The background map in the image is by the courtesy of OpenStreetMap.

**SPATIAL K-FOLD CROSS VALIDATION**

SKCV is a modification of the standard CV to overcome the biased prediction performance estimates of the model due to SAC of the data. The over-optimistic bias in the performance estimates is prevented by making sure that the training data set only contains data points that are at least a certain spatial or temporal distance away from the test data set.



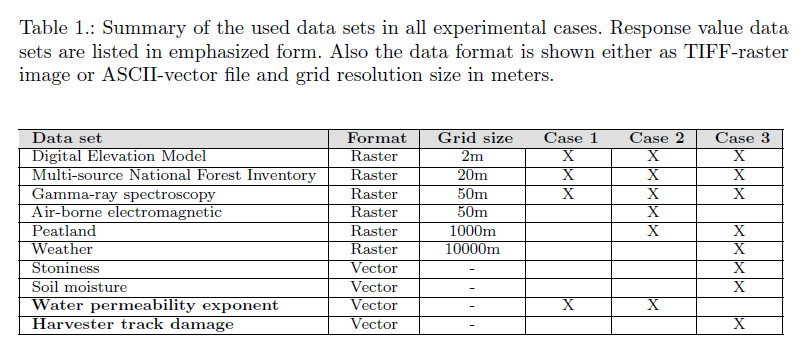


Table.1. Summary of the used data sets in all experimental cases. Response value data

sets are listed in emphasized form. Also, the data format is shown either as TIFF-raster

image or ASCII-vector \_le and grid resolution size in meters

**CONCLUSION**

Spatial-temporal autocorrelation is always present with GIS-based data sets and needs to be accounted for in machine learning approaches. As discussed above, traditional model performance criteria such as the CV method omit the consideration of the effect of SAC in the performance estimations with natural data sets. To account for the SAC in GIS based data sets we demonstrated by the means of three experiments that the SKCV method can be used for estimating the prediction performance of spatial models without the optimistic bias due to SAC, while the ordinary CV can cause highly optimistically biased prediction performance estimates. We also showed that SKCV can be used as a data sampling density selection criterion for new research areas, which will result in

reduced costs for data collection.

1. TOWARDS POSITIVE UNLABELED LEARNING FOR PARALLEL DATA**:** A RANDOM FOREST FRAMEWORK

By: -

Chen Li1, 2 and Xue-Liang Hua3

Department of Biochemistry and Molecular Biology, Monash University, VIC 3800, Australia College of Information Engineering, Northwest A&F University, Yangling 712100, China

Faculty of Information Technology, Monash University, VIC 3800, Australia

Springer International Publishing Switzerland 2014

**OBJECTIVE:** Parallel computing techniques can greatly facilitate traditional data

mining algorithms to efficiently tackle learning tasks that are characterized by

high computational complexity and huge amounts of data, to meet the requirement of real-world applications. However, most of these techniques require fully labeled training sets, which is a challenging requirement to meet. To address this problem, we investigate widely used Positive and Unlabeled (PU) learning algorithms including PU information gain and a newly developed PU

Gini index combining with popular parallel computing framework - Random

Forest (RF), thereby enabling parallel data mining to learn from only positive

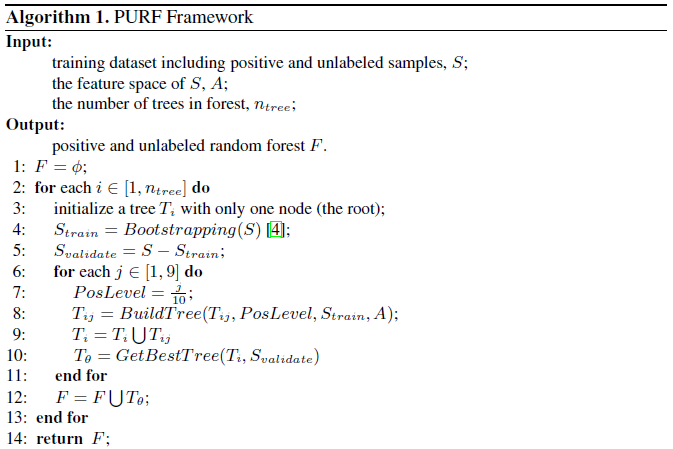
and unlabeled samples. The proposed framework, termed PURF (Positive Unlabeled Random Forest), can learn from positive and unlabeled instances

and achieve comparable classification performance with RF trained by fully labelled data through parallel computing according to experiments on both synthetic and real-world UCI datasets. PURF is a promising framework that facilitates PU learning in parallel data mining and is anticipated to be useful framework in many real-world parallel computing applications with huge amounts of unlabeled data.

**POSITIVE UNLABELED RANDOM FOREST FRAMEWORK**

In this section, the developed framework based on RF to learn from positive and unlabeled samples is described. PURF uses bootstrapping (with replacement) to select the training samples for each tree in forest. Therefore, after bootstrapping, approximately 2/3 of samples are collected as the training data. In each PU tree *Ti* in forest, there are nine sub-trees built in accordance with different values of *PosLevel* (ranging from 0.1 to 0.9 with a step size of 0.1). Then the rest 1/3 of the data will be used to determine the final output of each tree *Ti* in forest from the nine subtrees. Algorithm 1 shows the framework of the proposed PURF. Step 1 is used to initiate the forest *F*, *F* = *{T*1*, T*2*, . . . , Tntree}*. From steps 2 to 13, a forest *F* with *ntree* trees will be generated and returned. As mentioned above, each tree *Ti*(1 *≤* *i ≤ ntree*) in *F* has nine sub-trees with different *PosLevel* values. The function

*BuildTree*(*Tij, PosLevel, Strain,A*)(1 *≤ j ≤* 9) is used to generate sub-trees for *Ti*.



**CONCLUSIONS AND FUTURE WORK**

In this paper, for application in the parallel computing scenario, a novel framework termed PURF (Positive unlabeled Random Forest) with PU learning techniques that enables parallel data mining to learn from only positive and unlabeled samples was developed. Using the designed novel Gini index for PU learning as well as PU information gain, PURF-GI and PURF-IG with the PU Gini index and PU information based on RF, respectively is generated. Empirical assessment on real-world UCI datasets has strongly indicated that even provided with a high percentage of unlabeled data, PURFGI and PURF-IG were able to achieve an acceptable performance compared to RF. Our results on both synthetic and real-world dataset also indicate that PURF is powerful in learning from positive and unlabeled data. Experiments on real-world datasets showed that even with 90% unlabeled data, both PURF-GI and PURF-IG had a strong ability to distinguish positive from non-positive data. We also evaluated the performance and consumed time for calculation of PURF-GI in both parallel and non-parallel situations.

As expected, time analysis demonstrated that parallel PURF could indeed save considerable time through running multiple CPU cores and jobs. It is our expectation that this new framework will be increasingly used as a powerful approach to facilitate the processing and learning of positive and unlabeled data in the future. Therefore, in this research, we did not consider nominal attributes in our training datasets for simplicity, which is a limitation of this work. In the future, based on PURF-GI, we will investigate methods to deal with nominal attributes, thereby making it more suitable for real-world applications.

1. THE ADEPT K-NEAREST NEIGHBOUR ALGORITHM - AN OPTIMIZATION TO THE CONVENTIONAL K-NEAREST NEIGHBOUR ALGORITHM

By:-

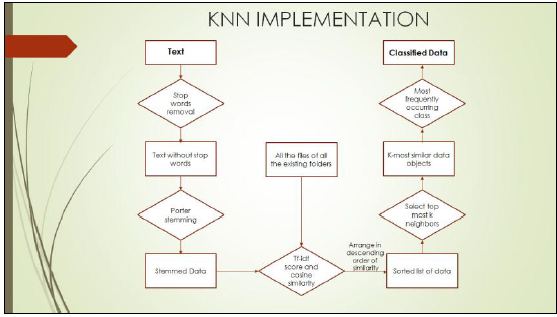
**Anjali Jivani1, Karishma Shah2, Shireen Koul3 and Vidhi Naik4**

Department of Computer Science and Engineering, The M. S. University of Baroda, Vadodara, India*.*

Journal on Machine Learning and Artificial Intelligence, Volume 4 No 1 February (2016)

**OBJECTIVE :** This research aims to study the efficiency of a well-known classification algorithm, K-Nearest Neighbour, and suggest a new classification method, an optimised version than one of the existing classification method. The purpose of this research is to reduce the time taken by the existing K- Nearest Neighbour Classification method. The classification algorithm’s purpose is to identify the characteristics that indicate the class to which each document belongs. This pattern not only helps in understanding the existing data but also to predict how new instances will behave. Classification algorithms create classification models by examining already classified data (cases) and inductively finding a predictive pattern.

**K-NEAREST NEIGHBOUR ALGORITHM (KNN):** kNN is an algorithm that classifies Sobservations such as euclidean, cosine, etc. kNN algorithm is a lazy learner i.e. it does not learn anything from the training tuples and simply uses the training data itself for classification. It is a non-parametric method used for classification and regression.

 Figure.1.Flowchart showing stepwise implementation of kNN

**THE ADEPT K NEAREST NEIGHBOUR ALGORITHM:** The newly suggested method, Adept kNN, emphasises on the fact that most of the comparisons or similarity checks go in vain. It is suggested that in contrast of comparing a document with each existing or previously classified document, it will be better to compare the document to be classified with three cases only (or files as in our example) each document being an intermediate representation of corresponding class. This intermediate representative document or text file is subjected to change over time. Adept kNN is just an optimisation of kNN.

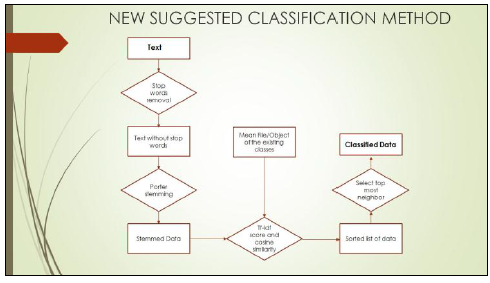


Figure.2. Flowchart showing stepwise implementation of Adept kNN

**FUTURE SCOPE**: The area of application of Adept kNN is almost similar to that of the original kNN .

1. Text Mining

2. Agricultural Area

3. Financial Departments

4. Medicinal Field

5. Large Organizations

1. **ON THE EVOLUTIONARY WEIGHTING OF NEIGHBOURS AND FEATURES IN THE K-NEAREST NEIGHBOUR RULE**

**By:-**

Daniel Mateos-Garc´ıa\_, Jorge Garc´ıa-Guti´errez, Jos´e C. Riquelme-Santos

Department of Computer Science, University of Sevi Avda. Reina Mercedes S/N, 41012 Seville (Spain)

Published date : September 12, 2017

**OBJECTIVE:** This paper presents an evolutionary method for modifying the behaviour of the k-Nearest-Neighbour classifier (kNN) called Simultaneous Weighting of Attributes and Neighbours (SWAN). Unlike other weighting methods, SWAN presents the ability of adjusting the contribution of the neighbours and the

significance of the features of the data. The optimization process focuses on

the search of two real-valued vectors. One of them represents the votes of

neighbours, and the other one represents the weight of each feature.

The synergy between the two sets of weights found in the optimization

process helps to improve significantly, the classification accuracy. The results

on 35 datasets from the UCI repository suggest that SWAN statistically

outperforms the other weighted kNN methods.

**METHODOLOGY USED:** In this section, we describe our weighting optimization method called Simultaneous Weighting of Attributes and Neighbours (SWAN). The purpose of this work and how the weighting vectors from the learning process are used have been presented in subsection.

1. Purpose and functionality.

**-**Individual encoding.

**-**Crossover and mutation.

**-** Error function.

**-**Generational policy.

**CONCLUSION:** This work presented a method to improve the kNN rule. We unified two classic paradigms of weighting by evolutionary computation. On the one hand, we adjusted the contribution of every neighbour used in the classification step, but nevertheless, the significance of the data features was modified simultaneously in order to achieve a better result in the recognition of new instances. In spite of the complexity of the solution to optimize and the increase of the search space in comparison with single-vector based methods, the experiments showed a successful behaviour of our approach.

1. **ASYMPTOTIC OPTIMALITY & EFFICIENT COMPUTATION OF THE LEAVE-SUBJECT-OUT CROSS-VALID5ATION**

By:-

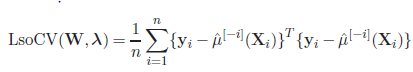
Ganggang Xu1 and Jianhua Z. Huang1,2

Texas A&M University

Published by the Institute of Mathematical Statistics in the journal of The Annals of Statistics,2012, Vol. 40, No. 6, 3003–3030.

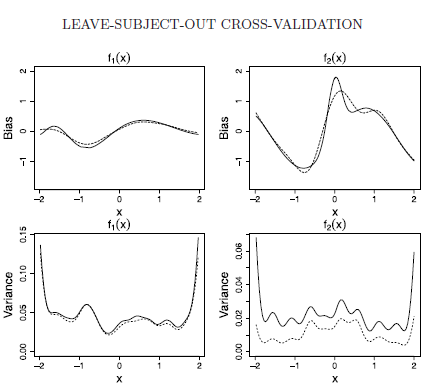
**OBJECTIVE:** Although the leave-subject-out cross-validation (CV) has been widely used in practice for tuning parameter selection for various nonparametric and semiparametric models of longitudinal data, its theoretical property is unknown and solving the associated optimization problem is computationally expensive, especially when there are multiple tuning parameters. In this paper, by focusing on the penalized spline method, we show that the leave-subject-out CV is optimal in the sense that it is asymptotically equivalent to the empirical squared error loss function minimization. An efficient Newton-type algorithm is developed to compute the penalty parameters that optimize the CV criterion. Simulated and real data are used to demonstrate the effectiveness of the leave-subject-out CV in selecting both the penalty parameters and the working correlation matrix.

**LEAVE-SUBJECT-OUT CROSS VALIDATION:** Let ˆμ(·) denote the estimate of the mean function obtained by using basis expansion of unknown functions fk’s (k = 1, . . . ,m) and solving the minimization problem (2) for \_. Let ˆμ[−i](·) be the estimate of the mean function μ(·) by the same method but using all the data except observations from subject i, 1 ≤ i ≤ n. The LsoCV criterion is defined as



By leaving out all observations from the same subject, the within-subject

Correlation is preserved in lsocv.



**CONCLUSION:** An approximation of leave-subject-out CV. A close inspection of the short-cut formula of LsoCV(W,\_) given in (16) suggests that the evaluation of LsoCV(W,\_) can still be computationally expensive because of the requirement of matrix inversion and the formulation of the hat matrix A. To further reduce the computational cost, using Taylor’s expansion (Iii − Aii)−1 ≈ Iii + Aii, we obtain the following approximation of LsoCV(W,\_) :



where ˆei is the part of ˆe = (I −A)Y corresponding to subject i. The next theorem shows that this approximation is a good one in the sense that its minimization is asymptotically equivalent to the minimization of the true loss function.

**RESEARCH PAPER STUDY**

**SWOT ANALYSIS**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **KEY POINTS** | | **RESEARCH PAPERS** | | | | | | |
| **I** | **II** | **III** | | **IV** | **V** | **VI** |
| **MACHINE LEARNING & DATA MINING TECHNIQUES USED** | **i. Supervised learning** |  |  |  | |  |  |  |
| **Ii. Unsupervised learning** |  |  |  | |  |  |  |
| **Iii. Reinforcement learning** |  |  |  | |  |  |  |
| **METHODOLOGY USED** | **i. K-Fold Categorization** |  |  |  | |  |  |  |
| **ii. Spatial K- Fold Cross Validation** |  |  |  | |  |  |  |
| **iii. Random Forest Framework** |  |  |  | |  |  |  |
| **iv. K-Nearest Neighbour** |  |  |  | |  |  |  |
| **v. Adept K-Nearest Neighbour** |  |  |  | |  |  |  |
| **vi. Text Categorization** |  |  |  | |  |  |  |
| **vii. Simultaneous Weighting of Attributes & Neighbour (SWAN)** |  |  |  | |  |  |  |
| **viii. Leave-one-subject-out (LOSO)** |  |  |  | |  |  |  |
| **ix. Randomization** |  |  |  | |  |  |  |
| **x. Naïve Bayes** |  |  |  | |  |  |  |
| **xi. Feature Selection** |  |  |  | |  |  |  |
| **xii. Brute Force** |  |  |  | |  |  |  |
| **xiii. Tokenization** |  |  |  | |  |  |  |
| **xiv. Cross Validation** |  |  |  | |  |  |  |
| **OPTIMAL SOLUTION** | **i. Optimal** |  |  |  |  | |  |  |
| **ii. Not Optimal** |  |  |  |  | |  |  |
| **SOFTWARE**  **COMPLEXITY** | **i. Worst Complexity** |  |  |  | |  |  |  |
| **ii. Average Complexity** |  |  |  | |  |  |  |
| **iii. Best Complexity** |  |  |  | |  |  |  |
| **SCALABILITY** | **i. Scalable** |  |  |  | |  |  |  |
| **ii. Not Scalable** |  |  |  | |  |  |  |

**PROBLEM STATEMENT**

Processing of huge amounts of data and high-complexity computations are the two emerging major challenges faced by in the above six researches. In the above analysis we gained in-depth understanding of the activity classification system and data mining concepts, and provide recommendation based on observed a problem in finding an optimal solution for multiclass classification.

**OBJECTIVE :-**

We need to explore One-Vs-Rest algorithm for multiclass classification. This algorithm involves training a single classifier per class, with the samples of that class as positive samples and all other samples as negative samples. This strategy requires the base classifiers to produce a real-valued confidence score for its decision, rather than just a class label; discrete class labels alone can lead to ambiguities, where multiple classes are predicted for a single sample**.** One-Vs-Rest algorithm takes a binary classifier as a parameter. We tried SVM, Logistic Regression, and LDA as the binary classifiers. Out of which LDA gave the best accuracy but not as good as Random Forest and kNN algorithm. So, we should focuse on RF and kNN to build a robust classifier.

The parameters of the estimator used to apply these methods are optimized by cross-validated grid-search over a parameter grid. Using these techniques we will apply the following parameters for RF and k-NN:

● Random Forest

○ Number of Decision Tree created : 50-200

○ Algorithm to be used : InfoGain(entropy)

● k-NearestNeigbor

○ Value of k varied from 9 - 11

○ Distance to be measured : euclidean

○ uniform weight for all the points

Python’s scikit-learn package will be used for performing all of the above mentioned tasks. This package provides standard machine learning libraries that we can modify as per our need and use it for our dataset. In addition to this, we will be using numpy, pandas for data processing and matplotlib for visualization purpose.