ISSN 2278-3091
Volume 9 No.2, March - April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse140922020.pdf
https://doi.org/10.30534/ijatcse/2020/140922020

Web Description Logic Rule Generation And Other Machine Learning Algorithms – A Comparative Study

Revathi. S[1], Dr. M. Suriakala[2]
[1]Assistant Professor, Department of Computer Applications, A.M Jain College, Chennai, India, revathibals.s@gmail.com
[2]Assistant Professor, Department of Computer Applications, Government Arts College for Men, Chennai, India, suryasubash@gmail.com

ABSTRACT

There has been an observation that the Online social network has gained booming popularity with large number of users over a period of time. The reason behind this ceaseless growth is the trust gained by people worldwide. OSN offers an interactive platform wherein people can build social ties, post their views, emotions, images and comment on other posts too. OSN users are so engrossed in building social circle that they hardly pay attention to any security related issues. In order to emphasize on the security and privacy parameters, a Web Description Logic Rule Generation (WDLRG) algorithm has been recommended for identifying and evaluating vulnerable users as well as eliminating the malicious users from the online social networks. Based on the sharing threshold, the algorithm can detect vulnerable users and safeguard them from any potential threats. Unlike other ML (Machine Learning) algorithms the proposed algorithm is effective and exhibits high success rate. The research presents a comparison study of WDLRG in contrast to rest of the ML Techniques namely Decision tree and Bayesian network.

Key words : Classification, Decision Tree, Online Social Networking, Supervised Learning, Vulnerable User, WDLRG.

1. INTRODUCTION

Machine Learning signifies a scientific study of algorithms and statistical models employed by the computer systems for carrying out a particular task effectively based upon patterns and inference rather than explicit instructions. ML represents a branch of AI (Artificial Intelligence) that constructs a mathematical model of sample data referred to as "training data" for drawing predictions/ conclusions without performing the task on explicit commands. Implementation of Machine learning algorithms are in various domains such as network intruder detection, email filtering and computer vision where building of a specific algorithm is not possible. In addition, machine learning spots the light on the construction of techniques that can learn and make predictions on available data. In instance, applications such as detection of network intruders or email filtering, optical character recognition (OCR), and computer vision. [1] The current research proposes a Web Description Logic Rule Generation (WDLRG) algorithm for identifying and evaluating vulnerable users as well as eliminating the malicious users from the online social networks. That is, based upon the maximum shared posts or information vulnerable users and attackers can be identified and in case the threshold limit surpasses, the malicious users can be removed. By employing sparse profile information it's clearly elucidated that the method is capable enough for identifying privacy attackers yielding an accuracy rate of 97% with just 2.67% of false negative. The same is being compared to rest of the equivalent methods involving big dataset and huge profile data. The two main approaches to treat a malware is static analysis, and dynamic analysis which has been explained in "Practical malware analysis" [2]. The recommended mechanism aids through building different levels such as forming social network, data uploading, machine learning techniques, examining user activities, detection and removal of malicious user. The work has generated certain unique features that aids in detecting and segregating the malicious users from the actual ones. Through a comparative study merits and demerits of one method in contrast to the other can be projected. It can act as a base for carrying out further important research helping the researches coming up with novel algorithms for concerns or requirements that are not addressed.

Following is the classification of the paper: Section 2 present decision tree induction, Section 3 illustrates BN (Bayesian networks), Section 4 presents WDLR (Web Description Logic Rule) and section 5 includes the comparative analysis of the proposed algorithm along with the conclusion.

2. DECISION TREE INDUCTION

Using Decision tree data can be classified into discrete form by the means of tree structure algorithms [3]. Decision trees emphasizes on exposing structural information present within data. It's a technique of supervised machine learning which attempts to generate a decision tree from a group of class labeled training samples during the process of machine

learning [4]. Beginning of DT (Decision trees) is with training samples along with their linked class labels. By considering the feature value, the training set is partitioned repeatedly into a subset in a way that each subset is purer compared to the parent set data. Every internal node of a decision tree depicts a test on attribute (feature), each branch depicts the test result and the class label is depicted by a leaf node. Decision tree acts as a classifier for finding the class label of an unknown sample by traversing path from root to the leaf node that has the class label for the concerned sample [4][5]. The training data can be ideally divided by the root node feature of the tree. Various other measures for identifying the feature that best divides the training data includes Information gain, Gini index, Gain ratio, myopic measures evaluates each attribute independently, Relief algorithm, Chi square, C-SEP, G-statistics, MDL (Minimum Description Length) measure is least bias toward multi-valued attribute, Multivariate split based upon combination of attributes [4][3][5][6]. Apparently, none of the measure surpasses the others [4]. Depending upon the tree height, complexity of the decision tree increases. Hence the preference is given to measures generating tree with multi-way and favoring more balanced splits based upon the dataset. Strategies employed by the basic decision tree induction algorithm includes non-backtracking, greedy, top-down and recursive divide and conquer.

2.1 Decision Tree Induction Algorithm Explanation

1. Node N is created and named it as Root node
2. Return Node N as leaf node, with class C.
3. Return Node N as leaf node with most common class, with no features.
4. Select best feature, by applying feature selection.
5. If feature found, set as test feature with Node N.
6. For each test feature, sample is partitioned and subtree is grown.
7. Assume ai as set of tuples and vi as test feature
8. If tuples (ai) is zero, attach a leaf node with the most common class in samples
9. If tuple is non zero, attach node returned by Generate decision tree

Tree pruning helps in enhancing the algorithm's prediction and classification accuracy by reducing over-fitting. In DT algorithms over-fitting may cause misclassification error. Tree pruning is carried out in a bottom-up manner and is easier than the tree growth phase since the training data set is scanned just one time. Tree pruning includes 2 methods, the first being Pre-pruning, wherein the tree is prune by hindering its construction early on the basis of pre-specified threshold and the second being post-pruning, wherein the sub-tree is eliminated from the fully grown tree. There is lot of computation involved in Post pruning but it yields much reliable trees. Prune trees can confront the issue of repetition and replication which can be attended and resolved using the multi-variant split along with a combination of attribute [4].

Popular decision tree algorithm includes ID3, C4.5 and CART. ID3 which being extension of concept learning theory by E. B. Hunt, J. Martin and P.T. Stone [7]. This being a recursive procedure that employs divide and conquer approach supporting nominal attributes only. Information gain enables the selected attribute to split but doesn't yield accurate output in presence of high noise or details in the training data set. Resultant, an in-depth pre-processing of data is performed prior to construct a decision tree model with ID3 [4]. C4.5 is built by [8], and utilizes gain ratio to perform attribute selection for splitting. It's enhanced compared to ID3 since it can handle nominal and numerical attributes and can deal with missing and noisy data. In C4.5, pruning is performed by replacing the internal node with a leaf node which minimizes the error rate. The C4.5 generated classifier can be depicted in form of decision tree as well as in a well comprehensible rule set form. Though the demerit being that the rule set form engulfs huge volume of CPU time and memory.

Breiman recommends Classification and Regression Trees (CART) [9] that makes use of Gini index measure for attribute selection in splitting and involves tests that are binary in nature. CART prunes trees with the help of cost-complexity model wherein parameters are assessed using cross-validation [10]. Merits, Demerits and Research concerns pertaining to Decision Tree Induction are mentioned below [4][11][5][10].

2.2 Merits

- Decision Trees (DT) are straight forward and quick
- They don't need any parameter setting or domain knowledge and are capable of handling high dimensional data.
- They are well comprehensible or easy to understand
- They yield a good accuracy depending on the data available.
- There is a support for incremental learning.
- They are unvaried as a single feature is used at each internal node

2.3 Demerits

- Involves high training time since it needs one pass over the training tuples for each level of tree.
- Memory shortage in case of bulkier databases
- Division of the instance space is orthogonal to the axis of one variable and parallel to rest other axes. After partitioning all the resulting regions are hyper rectangles.
- Most DT (decision tree) algorithms degrades in performance with issues involving diagonal partitioning.
- For certain concepts DT can be quiet tedious in terms of representation because of the replication issue

- Attributes order in tree nodes can pose negative impact on performance.

2.4 Research Issues

- Is it possible that a complex decision tree be split into small collection of simple trees, which when voted results in the similar output as the complex tree?
- IS it possible to build a non-trivial tree-construction algorithm without being impacted by neglecting a single case?

3. BAYESIAN NETWORK

Bayesian Classifiers are statistical classifiers which aids in predicting the class membership probability. It implies the probability that the available sample is associated to a specific class. Bayesian belief networks tends to be graphical models, depicting association among the subset of attributes. When imposed on bulkier databases, Bayesian classifier yields high accuracy and speed [11] [12]. NB (Naïve Bayes) Classifier is a simple statistical Bayesian Classifier and is referred to as Naïve since it considers that all the variables contribute for classification and are mutually linked. Such an assumption is referred to as class conditional independence [11] which remains unfeasible for many of the datasets. But it does results in a simple prediction framework yielding satisfactory output in various practical scenarios. The Naïve Bayes (NB) Classifier is founded on Bayes Theorem [13][14].

The Bayes Theorem is $P\ H|X = P\ X|H\ (H)(X)$ (1)

H represents some hypothesis, in a way that data tuple X belongs to a specified class C X – some evidence, describe by measure on set of attributes $PH|X$ – the posterior probability that the hypothesis H holds given the evidence X $P(H)$ – prior probability of H, independent on X $PX|H$ – the posterior probability that of X conditioned on H.

3.1 Merits

- It's simple to build and needs minimum computational time for training and very easy to construct.
- The model has a product type which can be transformed into a sum easily by the means of logarithms with remarkable consequent computational benefits
- It doesn't require any complex iterative parameter estimation technique hence can be imposed on huge datasets too.
- Easier interpretation of knowledge representation.
- It doesn't claim to be the superior most classifier but performs well enough and is robust.

3.2 Demerits

- In theory, Naïve Bayes (NB) classifier depicts low error rate in contrast to other classifier, but it doesn't hold true practically due to the assumption of class conditional independence and deprivation of available probability data.
- In accuracy it's inferior to rest of the classifier.

3.3 Research Issues

- What is the way to append extra edges for including certain dependencies amidst the attributes in order to resolve the assumption of attribute independence?

4. ESVMNN CLASSIFIER

Support Vector Machine (SVM) algorithm orients around the concept of a margin on either side of a hyper-plane, dividing two data classes. By increasing the margin and forming the largest possible distance amidst the separating hyper-plane and the instances on either side, the upper bound on the expected generalization error can be minimized [15][16].

The five feature sets that resulted after the implementation of feature reduction techniques were trained and tested by the means of improvised and newly built ESVMNN classification algorithm. Using the cross validation of 10 and 8-folds, performance of the classifier algorithm has been estimated.

ESVMNN Algorithm Explanation

1. SVM classifier identifies list of reduced features, named as S.
2. With 8 cross validation technique, divide data into training data, named as TrName and testing data named as TsName
3. Train the NN model with training data and SVM classifier.
4. Predict the output decision value as Xdecision, again train with decision value named as DLabel, set output as nnPredicted.
5. Calculate the classification accuracy.

5. GBoost

G Boost is one of the machine learning technique for the issue of regression and classification that generates a prediction model as an ensemble of weak prediction models or say literally of decision trees. Gradient Boosting is an ensemble learner like the Random Forests. In boosting, the standalone models are constructed sequentially by increasing the weight on instances with incorrect predictions and high errors rather than building on random subsets of data and features.

There is involvement of various training data sets wherein the values are provided accurately for efficient working of the algorithm. Gboost is based upon the principle that it builds a

strong rule from various weak learners. The said algorithm benefits by precisely predicting in the absence of any single utilized factors. Decision trees that are constructed are integrated to determine the predicted value which then helps in estimating the result. Following terminologies are utilized in this algorithm, namely pseudo residuals, shrinkage, decision trees, and prediction value.

GBoost Algorithm Explanation

1. Consider training set and differentiable loss function.
2. Set training model and calculate pseudo-residuals for each iterations
3. Train GBoost using training set and loss function
4. Compute multiplier and update model.
5. Output the value.

6. WDLRG ALGORITHM

It's assumed that users who have lots of friends have the tendency to expand their friend circle further. As a result priorities of such users may get impacted by the preferences of their explicit and implicit friends. The research imposes WDLRG (Web Description Logic Rule Generation) Algorithm for identifying and analyzing attackers and vulnerabilities. Emphasizes is laid on finding the most ideal unit set of aspects for identifying the vulnerable users in LinkedIn. The WDLRG algorithm aids in vulnerable user detection in accordance with the sharing threshold values and Extroversive Behaviors and eliminates the user that surpasses the set threshold limit.

WDLRG Algorithm Explanation

1. Consider a set of users S and threshold value.
2. Share threshold to every users.
3. Find extroversive and introversive behaviors
4. Identify user activity
5. Check any privacy vulnerable happens.
6. Check every user's sharing percentage per day.
7. Find vulnerable user, by taking sharing count. if count is greater than threshold, identify misactivity and remove user
8. Repeat for all users in the group.

6.1 Merits

- The proposed system effectively enables vulnerable user detection and application.
- It secures the user's account against any threat or hacking.
- The present research acts as a small input in the concerned area by safeguarding the users and reducing the amount of malicious activities.
- It yields output with highest accuracy.
- Minimum time

7. RESULTS AND DISCUSSION

Online social network has gained booming popularity with large number of users over a period of time. It offers the users an interactive platform to communicate and share any amount of data with anyone, anytime and anywhere. The interaction among the individuals involve sharing of their opinions and interests in form of text, images and videos.

OSN users are so engrossed in building social circle that they hardly pay attention to any security related issues. In order to emphasize on the security and privacy parameters, a Web Description Logic Rule Generation (WDLRG) algorithm has been recommended for identifying and evaluating vulnerable users as well as eliminating the malicious users from the online social networks. The recommended approach proves to yield improvised outcome pertaining to the OSN for vulnerable user detection when concerned with privacy. Attributes have been assembled from different researches and are spotted via intense assessment at the initial phase. The attributes are then weighted properly.

Table 1: Comparison Classification Techniques

S.No	Techniques	Accuracy (%)	Time (ms)
1	Bayesian networks	93.44	33:78
2	Decision tree	94.74	28.10
3	Enhanced SVM NN	95.18	24.31
4	GBoost	96.23	20.16
3.	Web Description Logic Rule Generation	97.77	17.12

Table 1, validates the comparison of performance for detecting vulnerable user profile processing employed in the classification methods to compare various machine learning systems such as the Decision tree, Bayesian networks, ESVMNN, GBoost with WDLRG (Web Description Logic Rule Generation). It's well elucidated in the research that the WDLRG yields improvised results in contrast to the existing methods.

The bar chart presents, the comparison of graph performance for detecting vulnerable user profile processing employed in the classification methods to compare various machine learning systems such as the Decision tree and Web Description Logic Rule Generation (WDLRG). It's well elucidated in the research that the WDLRG yields improvised results in contrast to the existing methods. Figure 2 depicts accuracy outcome of WDLRG

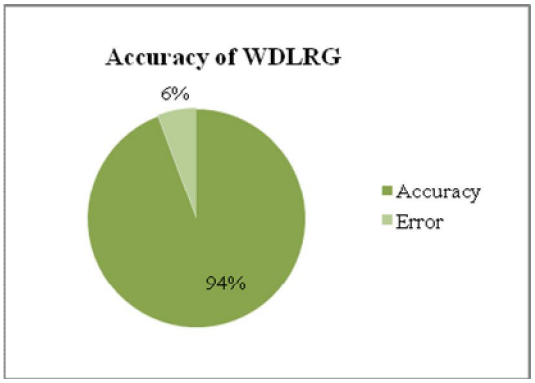

Figure 1: Accuracy Result

Figure 1 depicts that the error performance for identifying vulnerable user profile processing method on the online social network reveals an accuracy of 97.33% and error rate as 2.77%.

8. CONCLUSION AND FUTURE WORKS

The research has compared the popular classification algorithms such as the DT - decision trees, BN - Bayesian network, Gboost, ESVMNN in contrast with WDLRG. The motive behind the research was on comprehending their key ideas and identify existing research concerns that can contribute to the researchers and students who are carrying out advanced course on classification. The comparative study reveals that any given algorithm exhibits its own pros and cons along with its own space of implementation. Though it can be concluded that WDLRG tends to be more accurate compared to rest of the Classifiers. Though a classifier can be constructed by combining two or more classifier so as to add up their power.

REFERENCES

1. M. A. Ottam, N. A. Alawad and K. M. O. Nahar. **Classification of Mushroom Fungi Using Machine Learning Techniques**, *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 8, No. 5, pp. 2378 - 2385, 2019. https://doi.org/10.30534/ijatcse/2019/78852019
2. A. Amer and N. A. Aziz. **Malware Detection through Machine Learning Techniques**, *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 8, No. 5, pp. 2408 - 2413, 2019. https://doi.org/10.30534/ijatcse/2019/82852019
3. J. R. Quinlen. **Introduction of Decision Trees**, *Machine Learning*, Vol. 1, pp. 81-106, 2009.
4. J. Han and M. Kamber. **Data Mining Concepts and Techniques**, *Elsevier*, 2011.
5. S. B. Kotsiantis. **Supervised Machine Learning: A Review of Classification Techniques**, *Informatica*, Vol. 31, pp. 249-268, 2007.
6. T. N. Phyu. **Survey of Classification Techniques in Data Mining,** *International Multi-conference of Engineers and Computer Scientists*, 2009.
7. J. R. Quinlan. **C4.5: Programs for machine learning**," *Morgan Kaufmann*, 2012.
8. K. P. Soman. **Insight into Data Mining Theory and Practice,** *PHI*, 2006.
9. H. Jiawei. **Classification and Regression trees**, *Wadsworth, Belmont*, 2012.
10. J. R. Quinlan. **Discovering rules by induction from large collections of examples**. *Expert Systems in the Microelectronic age*, pp. 168 - 201, 2011.
11. T. Cover. **Nearest Neighbor Pattern Classification,** *IEEE Transactions on Information Theory*, Vol. 13, No. 1, pp. 21 - 27, 2011. https://doi.org/10.1109/TIT.1967.1053964
12. D. E. Rumelhart, G. E. Hinton and R. I. Williams. **Learning internal representation by error propagation**, *Parallel Distributed Processing*, 2014.
13. N. Friedman, D. Geiger and M. Goldazmidt. **Bayesian Network Classifiers,** *Machine Learning*, Vol. 29, pp. 131 - 163, 2010. https://doi.org/10.1023/A:1007465528199
14. Xindong Wu. **Top 10 algorithms in data mining**, *Knowledge Information system*, Vol. 14, pp. 1 - 37, 2014.
15. C. Cortes. **Support Vector Network**, *Machine Learning*, Vol. 20, pp. 273 - 297, 2015. https://doi.org/10.1007/BF00994018
16. R. Burbodge and B. Buxton. **An introduction to Support Vector Machines for Data Mining**, *Computer Science Dept., UCL, UK.*