Intelligent Association Classification Technique for Phishing Website Detection

Mustafa Al-Fayoumi[1], Jaber Alwidian[2], and Mohammad Abusaif[2]
[1]Computer Science Department, Princess Sumaya University for Technology, Jordan
[2]Big Data Department, Intrasoft Middle East, Jordan

Abstract: *Many critical applications need more accuracy and speed in the decision making process. Data mining scholars developed set of artificial automated tools to enhance the entire decisions based on type of application. Phishing is one of the most critical application needs for high accuracy and speed in decision making when a malicious webpage impersonates as legitimate webpage to acquire secret information from the user. In this paper, we proposed a new Association Classification (AC) algorithm as an artificial automated tool to increase the accuracy level of the classification process that aims to discover any malicious webpage. An Intelligent Association Classification (IAC) algorithm developed in this article by employing the Harmonic Mean measure instead of the support and confidence measure to solve the estimation problem in these measures and discovering hidden pattern not generated by the existing AC algorithms. Our algorithm compared with four well-known AC algorithm in terms of accuracy, F1, Precision, Recall and execution time. The experiments and the visualization process show that the IAC algorithm outperformed the others in all cases and emphasize on the importance of the general and specific rules in the classification process.*

Keywords: *Data mining, Association Classification technique, Apriori algorithm, Phishing.*

Received January 28, 2019; accepted March 28, 2019
https://doi.org/10.34028/iajit/17/4/7

1. Introduction

Data mining is a subfield of computer science that aims to discover hidden patterns or knowledge in large datasets, and to transform this extracted information into a more suitable form to enhance the decision-making process in many fields [6]. Data mining includes a set of techniques for different purposes, such as classification, clustering, regression, association rules, and Association Classification (AC) [1, 4, 19, 25, 27]. In this paper, we will shed light on the AC technique and its impact on enhancing the decision-making process in the phishing.

The understandability and simplicity of the rules that can be generated using the AC technique and its positive effect on the accuracy of the classification process or decision-making process make the AC technique attractive for researchers. However, the AC mining technique does have a disadvantage, namely that it generates a very large number of rules, requiring more memory and time than the classical data-mining techniques. Furthermore, many of the AC algorithms do not behave in a stable way in all datasets, so may not be scalable owing to this negative aspect [2, 12, 26, 28].

During the last decade, most of the government and financial organizations have explored their online services to their customers. In 2011, 85% of Europeans and 83% of Americans shopped online in regular form [20]. With the emerging use of smart phones, number of people increases based on number of online services to shop, pay their bills or check their banking account etc. While such services had a critical impact on the world economy, such as financial services that increase the level of security risks for both financial institutes and customers.

Phishing is a criminal technique uses two concepts, which are technical subterfuge and social engineering to steal financial account credential and consumer's personal identity data. Phishing consider as a new identity theft crime. Where, the media reports many of stories about an organization that has clients targeted by a phishing attack that makes the financial organizations protect their customers by improving their security techniques, phishers develop even more sophisticated attacking techniques [27].

Many of malicious people create phishing websites that are fake web pages to imitate web pages of real websites [20, 23, 27]. In order to scam victims phisher create webpages that are visually very similar to the real webpages. This kind of scam targeted an unaware client easily. The victims of a phishing Web page may expose their bank password, credit cards numbers, accounts, or other. In addition, phishing is a relatively new electronic crime when compared to other forms (e.g.,) hacking and viruses [3].

The main aim of this paper is to build a more accurate intelligent AC technique and apply it to a phishing website of UCI dataset. A comprehensive experimental study using UCI dataset will be presented to evaluate and compare well-known association rule-based classification techniques with our proposed technique in terms of accuracy, F1, recall, precision, and building time for the model. Furthermore, our study aims to meet the following objectives:

- Conduct a comprehensive and significant study on some aspects of AC data-mining techniques.
- Develop scalable and accurate AC algorithms.
- Produce extensive experimental results to evaluate the proposed AC technique with regard to different datasets from the UCI repositories.

The main of the AC concepts are presented in section 2. In section 3, some relevant works are discussed. We describe our proposed model in detail in section 4. Extensive experimental results are given in section 5. Finally, the conclusion and discussion of future work are presented in section 6.

2. AC Background

The AC technique merges the classification and association rules concepts in one form. The association rules algorithms developed to discover a hidden correlation between attributes, while the classifiers developed to predict the class label. Thus, the AC technique is considered as the second generation for both techniques, and is designed to find the hidden correlation between attributes and classes. For example, in a rule such as attribute1 (V1), attribute2 (V1) → Class (V1), Class (V1) must be a class value, while attribute1 (V1) and attribute2 (V1) are attribute values. This rule can be interpreted as meaning that if attribute1 (V1) and attribute2 (V1) attribute values occur together for any object; this object can be classified as Class (V1), which represents the class value [1, 2, 18, 22].

The Classification Based on Association Rules (CBA) algorithm was proposed in [22] to employ the association rules in the classification task that produced new generation in the classification process that's called AC technique. This algorithm was built on three phases: rule generation, pruning, and prediction. In the rule generation phase, Apriori algorithm was used to generate the frequent itemset that represent the Class Association Rules (CARs) where, all of these frequent itemset should pass two estimated measures (minimum support and minimum confidence), as shown in following steps:

1. Generate the candidate single itemset. And then generate the frequent single itemset, based on selecting the items that have support greater than or equal to the estimated minimum support. The Support for any item can be calculated by Equation (1).

$$Support = \frac{(X \cup Y).count}{n} \tag{1}$$

Where, x is the attribute, y the name of the class, and n the number of rows in the dataset.

2. Generate the candidate 2-itemset.
3. Generate the frequent 2-itemset that passes the minimum support.
4. Repeat to find all next itemsets until the set is empty.
5. Finally, The CARs should be generated from the frequent sets by selecting the rules that have confidence greater than or equal to the estimated minimum confidence, where the confidence of an item can be calculated by Equation (2).

$$Support = \frac{(X \cup Y).count}{X.count} \tag{2}$$

After finding the CARs using the Apriori algorithm, the M1 method is used in the pruning phase to select the best rules that cover the entire database. Finally, the prediction phase predicts the class for any given unknown input, the class of the first rule that can match this input will be assigned as its predicted class.

3. Related Works

Phishing has a negative impact on many field inside the organizations such as relationships of customer, revenues, efforts of marketing, and overall corporate image. Phishing attacks may cost organizations millions of dollars for each attack in personnel time and fraud-related losses. The worst, costs related with the damage to consumer confidence can run in the millions of dollars [10, 11, 13, 21].

The wide variety of data being captured leads to make quick retrieval of information and efficient management very critical issue for the decision making process [14]. Data mining is the science of deriving meaningful knowledge from large data sets [29]. Knowledge discovery techniques have been employed in several areas such as decision support, market analysis, financial analysis, and industrial retail [7, 15].

Various experimental studies have found that AC techniques perform better than the traditional classifiers owing to the small number of rules that can be produced by the traditional techniques. Moreover, AC techniques produce a large number of important rules that cannot be generated by traditional classifiers, and these rules can enhance the entire classification process [1, 5, 7, 19, 22, 27].

Classification Based on Multiple Association Rules (CMAR) algorithm was developed by Li *et al*. [17]. The MCAR algorithm depends on the association rules and the classification techniques. This algorithm developed a new approach to generate rules and proposed a new phases for the classifier. The author use CR-tree and FP-tree algorithms to generate hidden rules instead of using the original Apriori. The classifier works on finding the class value for its unknown instances by generating all hidden rules that could be used with these instances and then evaluate all of these hidden patterns for prediction. CMAR was compared with set of algorithms in terms of statistical measures using UCI datasets. The experimental study showed that the CMAR algorithm outperformed the other compared algorithms.

Hadi *et al*. [13] developed a new AC algorithm that is called new Fast Associative Classification Algorithm (FACA) [13]. The author used Diffset method to generate rules that leads to enhance the speed of building the model. Furthermore, the selected rules are sorted according to set of measures: the least number of attributes, confidence, support and first occurrence respectively. FACA also developed a multiple rules approach for the prediction phase to increase the accuracy level of the classifier. In particular, FACA algorithm divides the rules that match the given instance to set groups based on the label and then selects the label of the largest group. For the evaluation process, the authors compared their algorithm with some of AC algorithms based on set of statistical measures.

The Enhanced CBA (ECBA) algorithm was developed by using a new Statistical Measure by Alwidian *et al*. [9]. The ECBA algorithm was evaluated in terms of accuracy and building time. Experimental results showed the ECBA was more accurate than the other algorithms and give acceptable building time model in most of experiments. Furthermore, the selected statistical ranking measure that is used in the rule generation process does not solve the problem of the estimate measure where, some of rules will not be generated if the support of these rules has value less than the minimum support value with very small difference.

The Apriori algorithm was optimized using general rule generation by Alwidian *et al.* [8] to overcome the long time needed in the generation phase to achieve incremental application. The authors proposed the FCBA algorithm and compared this with a set of AC algorithms in terms of accuracy, recall, precision, F1, and building time model measures.

The main critical issue that could be observed in all of these algorithms is the type of rules in the rule generation process. Some of these algorithms depends on specific rules while the others depend on general rules and this can lead to miss very important rules to enhance the classifier. In our proposed solution will make combination between of these types to enhance the classification process.

The explanation for three critical issues related to the phishing detection are presented in [16]. The author in this article, describes how the hackers could hack the web pages, how to prevent themselves and how they are threat to E-business. In addition, evaluate the Multi-label Classifier based Associative Classification (MCAC) as an AC algorithm and show how the AC techniques could be employed in the phishing detection.

Sankhyan *et al.* [24], focuses on the process of feature selection and how can select subset of features to enhance the classification process in the phishing detection. Aim of this paper is to give accurate classifier to predict the URLs that are legitimated and phishing.

4. Proposed Model

We proposed an intelligent AC algorithm to enhance the phishing detection process. The IAC algorithm aims to generate general and specific rules within the rule generation phase. In addition to enhance the building time model and the accuracy level of the classifier.

The IAC algorithm employs the Harmonic Mean (HM) measure in all phases of the AC technique that leads to overcome the weakness of estimated measures (support and confidence). Furthermore, employing the HM measure in the pruning and prediction phases in our algorithm increased the accuracy level of the classifier, while the building time model enhanced due to use the HM measure in the rule generation process. The IAC algorithms depends on the following assumption:

- Assumption: The IAC algorithm prefers the general and specific rules to cover large size of the dataset. In other words, if have general rule like X→T and this rules exceeds the minimum HM measure then, this rule will be generated directly as final rule in our classifier without need to generate the next generations that could be derived from the same rule that leads to enhance the building time model and accuracy of the classifier. Using the HM measure solves the problems of the estimated measures by creating one harmonic value to represent the support and confidence and use it in all phases.

Furthermore, the IAC algorithm generates specific rules with high confidence, that's mean the rules have support value less than the minimum support value but on other side, it has very high confidence value. The specific rules will represent the mutations in the dataset which may be very important for many critical applications, precisely in the phishing detection process, where the phishing could be consider as strange cased to be discovered.

4.1. The IAC Algorithm Structure

The IAC algorithm designed based on four main phases: Pre-processing, Rule generation, Pruning and Predication phases as shown in Figure 1. In our algorithm, we generate minimum HM value before all of these phases based on the minimum support and minimum confidence, Equation (3) used to generate the minimum HM value while Equation (4) used to generate the HM value for each rule.

Minimum HM (Minimum Support, Minimum Confidence) =

$$\frac{2}{\frac{1}{Minimum\ Support}+\frac{1}{Minimum\ Confidence}} \qquad (3)$$

HM (Support (r), Confidence (r)) =

$$\frac{2}{\frac{1}{Support(r)}+\frac{1}{Confidence(r)}} \qquad (4)$$

4.2. The IAC Algorithm Description

Algorithm1 shows the workflow of the IAC algorithm, which involves four phases:

1. Pre-processing.
2. Rules generation.
3. Pruning.
4. Prediction.

Algorithm 1: IAC Algorithm (T, n)

```
Dataset T with n training objects
IAC (T, n)
{
Divide (T) →[Training-data, Test-data]
S = empty set
Compute_ Minimum_ Harmonic_ Mean (Minimum support, Minimum Confidence)
S = Generate-Rule (Training-data, n, Minimum Harmonic Mean)
Rule-Pruning (S, Minimum Harmonic Mean) →[strong-rules]
Prediction (best rules, Test-data)
}
```

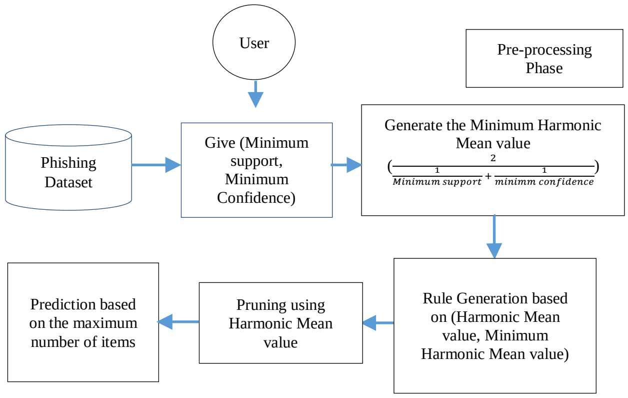
User
Pre-processing Phase
Phishing Dataset
Give (Minimum support, Minimum Confidence)
Generate the Minimum Harmonic Mean value
$\left(\frac{2}{\frac{1}{Minimum\ support}+\frac{1}{minimm\ confidence}}\right)$
Prediction based on the maximum number of items
Pruning using Harmonic Mean value
Rule Generation based on (Harmonic Mean value, Minimum Harmonic Mean value)

Figure 1. The IAC phases.

After generating the minimum HM value in the pre-processing phase from the estimated measures that are given by the user by using Equation (3) , the IAC algorithm begin with the rule generation phase. During this phase, our algorithms compute the HM value for each single item by using Equation (4) and keep those rules in the Class Association Rules (CARs) that have HM value greater than or equal the minimum HM measure. All of the rules that satisfy the previous condition is inserted in the CARs while the other is used to span the next generation of item rules. This process will be repeated to evaluate all generations and select the optimal association rules to build the classifier as shown in Algorithm 2.

In the pruning phase, the IAC algorithm sorts all rules based on the HM value form the highest to the lowest and uses the M1 method for database coverage to eliminate unneeded rules and keep the remaining as rules of the classifier to be used in the prediction phase as shown in algorithm 3. The predication phase uses the maximum number of items in the rule to decide, which rule is winner (i.e. let we have this instance (X, Y, Z→?), and should be classified. By using our classifier we found two rules could classify this unknown instance: X→T and X, Z →F, our classifier will select the second rule which is X, Y→F because of it is more specific that's mean contain maximum number of items on the left side). On other scenario, if the rules that match the given unknown instance have the same number of items on the left side the rule that has maximum HM value will be selected and if more than rules have the same HM value then the first occurrence technique employs.

5. Experimental Results

We performed an extensive analysis to assess the accuracy, F1, precision, recall, and building time model. The IAC algorithm was compared with four well-known AC algorithms-CMAR, FACA, ECBA and FCBA - based on a set of experimental results. We conducted our experiments on a 3GHz i3 PC with a 4GB main memory. The IAC algorithm is implemented using the Java programming language within the WEKA tool. While, the compared algorithms were implemented by the authors. The parameters of these algorithms were three pairs of minimum support and minimum confidence values as follows: (0.2, 0.5), (0.1, 0.4) and (0.05, 0.3).

Algorithm 2: Rule generation phase (T, minimum support, minimum confidence)

```
Compute the minimum HM
$Minimum\ HM\ (miinimum\ support, minimum\ confidence) = \frac{2}{\frac{1}{Minimum\ support}+\frac{1}{minimm\ confidence}}$
Find the HM value for each candidate rule (r)
    $HM\ (support(r), confidence(r)) = \frac{2}{\frac{1}{support(r)}+\frac{1}{confidence(r)}}$
For each candidate rule (r)
(
If (HM >= minimum HM)
Then
        Add to CARs
else
        Leave it in the Itemset
If (all candidate rules tested)
Then
        Generate the Itemset S'
        S=S'
If S is not empty
Then
        Go to step 4
End
)
```

```
Algorithm 3: Pruning (S, Minimum Harmonic Mean)

{
Sort (CAR) # Sort the rules based on the Harmonic Mean
values descendingly.
M1 (CAR) →[best rules] # (Liu et al., 1998)
Return (best rules)
}
```

5.1. Dataset

We used phishing dataset from the UCI repository that contains 1353 instances and 10 attributes. Author of the dataset identified set of features related to legitimate and phishy websites and gathered 1353 instances, where each instances represents different website from difference sources. These instances were collected from Phishtank data archive (www.phishtank.com), which is a free community site gives the users set of authorities such as submit, track, share phishing data and verify. The legitimate websites were gathered from Yahoo and starting point directories using a web script developed in PHP. The PHP script was plugged with a browser and the author gathered 548 legitimate instances out of 1353 instances. There is 702 phishing instances, and 103 suspicious instances. Name of features, number of distinct values in each feature and their labels are shown in Table 1.

Table 1. Phishing dataset features.

Name of Attribute	Number of Distinct Values	Labels
URL Anchor	3	-1,0 and 1
Request URL	3	-1, 0 and 1
SFH	3	1,-1 and 0
URL Length	3	1,-1 and 0
Having_IP_address	2	0 and 1
Popupwindow	3	-1, 0 and 1
SSL_final_state	3	1,-1 and 0
Web traffic	3	1, 0 and -1
Age of domain	2	1 and -1
Result	3	0,1 and -1

5.2. Analysis of Results

To evaluate the performance of the considered algorithms, we used set of statistical measures that are *F1*, Precision, and Recall. Where, *F1* measure is calculated using Equation (5).

$$F1 = \frac{2*(Precision*Recall)}{Precision+Recall} \tag{5}$$

As a machine learning statistical measures, we used *Precision* and *Recall* measures in the evaluation process that are calculated by using Equations (6),and (7), according to Table 2.

Table 2. Confusion matrix for classes.

Name of Attribute	Predicted as	
	Actual Class	**Other Classes**
Actual Class	**True Positive (TP)**	**False Negative (FN)**
Other Classes	**False Positive (FP)**	**True Negative (TN)**

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

Table 3, presents a comparison of the Accuracy, F1, Precision, Recall and building time model of CMAR, FACA, ECBA, FCBA and IAC algorithms where minimum support=0.2 and minimum confidence=0.5. In this experiment, the IAC algorithm outperformed the others algorithms in terms of all measures with accuracy 81.8921 % due to the variations in the rule generated types. Furthermore, the ECBA algorithm comes in the second place with accuracy 78.7879 %. While, the FCBA algorithm comes in the last place with accuracy 55.497 % due to the internal process inside this algorithm that depend only on the general rules. The most observed issue in this experiment is the building time model measure, where the values in this measure are closed for each other due to the small size of the dataset.

Table 3. Evaluation of CMAR, FACA, ECBA, FCBA and IAC algorithms based on Accuracy, F1, Precision, Recall and building time model measures, with minimum support =0.2 and minimum confidence.

Algorithm name	Accuracy	F1	Precision	Recall	Building Time Model
CMAR	70.1404	0.659	0.690	0.701	0.21 s
FACA	70.1404	0.659	0.690	0.701	0.19 s
ECBA	78.7879	0.753	0.738	0.788	0.37
FCBA	55.9497	0.756	0.726	0.789	0.13 s
IAC	81.8921	0.795	0.791	0.819	0 s

With minimum support and minimum confidence values 0.1 and 0.4 respectively, the second experiment are executed to emphasize the IAC algorithm outperformed all considered algorithms in terms of all measures. Where, the accuracy of the IAC algorithm was better than the accuracy in the previous experiment as shown in Table 4. The same scenario was repeated for the second and last places.

Table 4. Evaluation of CMAR, FACA, ECBA, FCBA and IAC algorithms based on Accuracy, F1, Precision, Recall and building time model measures, with minimum support =0.1 and minimum confidence.

Algorithm name	Accuracy	F1	Precision	Recall	Building Time Model
CMAR	78.9357	0.755	0.736	0.789	0.79 s
FACA	78.4183	0.750	0.733	0.784	0.67 s
ECBA	82.7051	0.794	0.805	0.827	1.49 s
FCBA	69.9187	0.769	0.737	0.803	0.1 s
IAC	83.8137	0.835	0.832	0.839	0 s

The best performance on most of these algorithm was in the third experiment because of the values of minimum support and minimum confidence are reduced to generate more association rules in the classifier that leads to enhance the accuracy level for most of these algorithms as shown in Table 5. The IAC algorithm came in the first place with accuracy 85.6359% and the ECBA algorithm came in the second place with accuracy 84.1833%. The FCBA algorithms moves from the last place to get the third one with accuracy 82.705%. Furthermore, for extensive analysis we compared all of these algorithms in tem of average accuracy for all experiments to show the overall behaviour as shown on Figure 2.

Table 5. Evaluation of CMAR, FACA, ECBA, FCBA and IAC algorithms based on Accuracy, F1, Precision, Recall and building time model measures, with minimum support =0.05 and minimum confidence = 0.3.

Algorithm name	**Accuracy**	**F1**	**Precision**	**Recall**	**Building Time Model**
CMAR	76.4228	0.730	0.718	0.764	0.27 s
FACA	51.8847	0.354	0.269	0.519	0.15 s
ECBA	84.1833	0.811	0.809	0.842	8.92
FCBA	82.7051	0.794	0.805	0.827	1.5 s
IAC	85.3659	0.857	0.858	0.857	0.01 s

Moreover, we visualized the data distribution based on two criteria to validate our assumption. The first criteria, visualize the data based on single feature to show if there is general rules could be generated in the classifier, where these rules should have high accurate indication for the class value. The second criteria depend on multi feature rules that called specific rules. As shown in Figures 3, 4, 5, and 6.

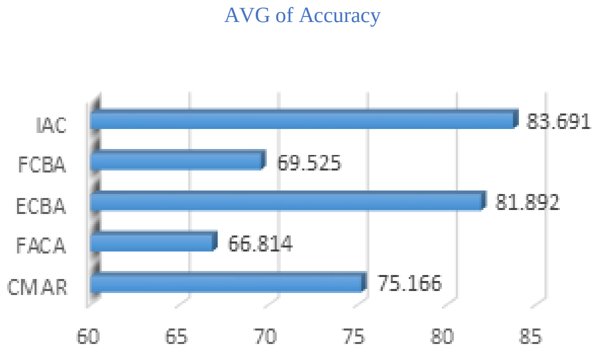

Figure 2. average of accuracy in all experiments.

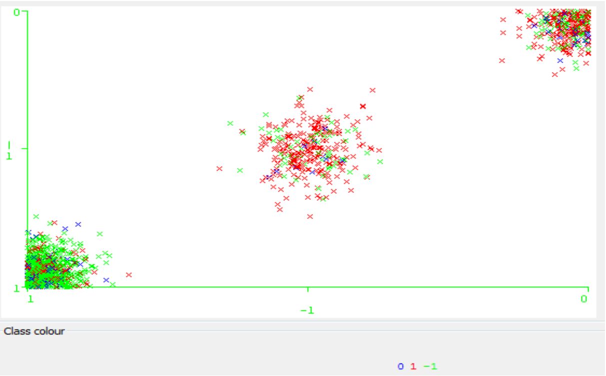

Figure 3. SSL_final_state feature Distribution on the class type.

Figure 3 presents the distribution of the SSL_final_state feature based on type of class. The SSL_final_state feature contains three distinct values with three labels: 1, -1, and 0. while, the class has also three distinct class values with labels: 0, 1 and -1. Class with label 0 represented in the figure by using blue colour, class (1) by using red colour and finally, class (-1) represented by green colour.

It is noticeable that most of the SSL_final_state values with label 1 have class with label (-1) and that is clear from the green colour on the lower left corner in the Figure. This observation suggests that the single rule that could be generated from this feature when the label is 1 can give accurate indication about the class with label -1 (i.e. SSL_final_state (1) →class(-1)).

On other words, Figure 3, shows the importance of the general rules approach in the classification process. The scenario is repeated in Figure 4 that represents the distribution of the Request_URL feature that has regular behavior regarding to the value of label 1 on the upper right corner of the figure with class (-1) that has green color (i.e, Request_URL (1)→ class (-1)). This conclusion, leads to eliminate the features that have irregular behavior and that very important in the prediction phase.

To show the importance of the specific rules in the prediction process, we visualize two attributes with the class value. Figure 5, represents Request_URL as (Y-axis) and URL Anchor as (X-axis) while the class value represents color of the point. Furthermore, each attribute has three labels that means, each attribute contains three distinct values. In the analysis of this Figure, we find that the intersection between Request_URL with label 1 and URL Anchor with label 0 holds class 1 for most of instances that are represented in red color, where number of these instances is small if compared with number of instances in the dataset. This makes us able to conclude the following: Request_URL (1), URL Anchor (0) → class (1) and this rule has support less than the minimum support but at the same time it has very high confidence value and it represents high confident rule in the classifier.

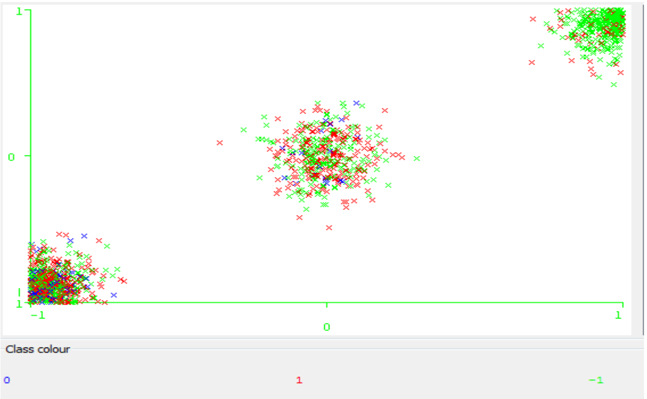

Figure 4. Request_URL feature Distribution on the class type.

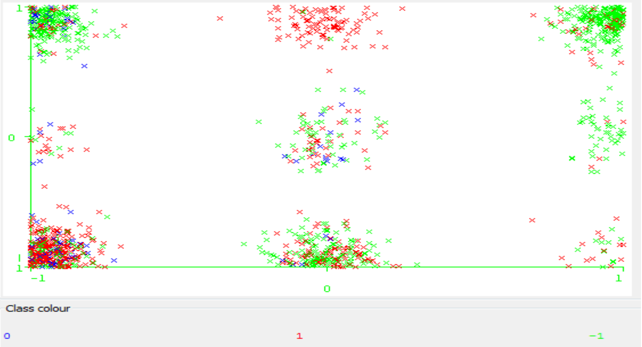

Figure 5. Request_URL as (Y-axis) and URL Anchor as (X-axis) features Distribution on the class type.

In addition, Figure 6 shows different specific rules for Web_Traffic as (Y-axis) and SSL_final_state as (X-axis) features. In this Figure, we can generate three rules based on the predominant green color at the bottom of the figure as follows:

Web_Traffic (1), SSL_final_state (1) → Class (-1)
Web_Traffic (1), SSL_final_state (0) → Class (-1)
Web_Traffic (1), SSL_final_state (-1) → Class (-1)

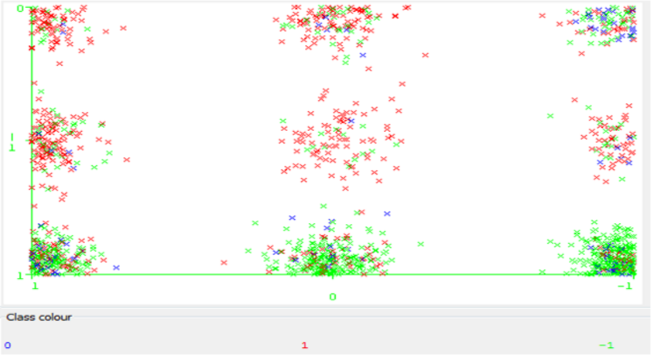

Figure 6. Web_Traffic as (Y-axis) and SSL_final_state as (X-axis) features Distribution on the class type.

Finally, the visualization process could eliminate many of features that do not reflect any relations between of them as shown in Figure 7. This Figure shows how there is no relation between URL-length and age-of-domain features based on class feature.

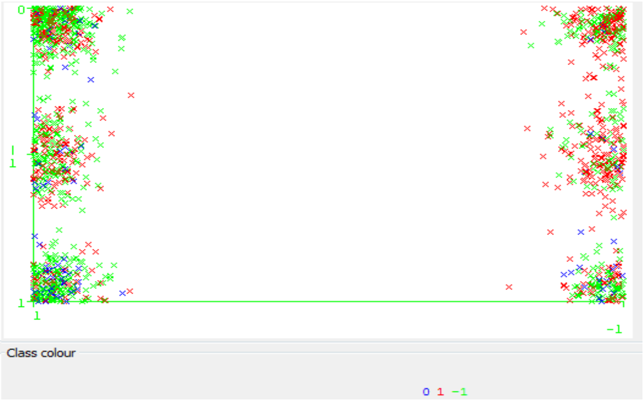

Figure 7. URL-length as (Y-axis) and age-of-domain as (X-axis) features Distribution on the class type.

6. Conclusions

In this paper, we investigated a new association classification algorithm named IAC. The IAC algorithm aimed to enhance the accuracy of classifiers based on an efficient statistical measure. Aim of this algorithm is to solve the problems of the estimated measures that are playing main role in all AC techniques to discover rules. Our new algorithm was developed based on a Harmonic Mexan Measure in all phases of the association classification technique to generate rules that are more useful. All these features leads to develop new pruning and predication approaches that improve the performance of the IAC algorithm in terms of several measures. We have tested the IAC algorithm against four modern AC algorithms running on phishing dataset as a case study. The experiments show outperformed behaviour for the IAC algorithm. Furthermore, the visualization approach for the distribution of the data emphasized the powerful of the HM measure in the association classification technique and as future works will investigate several statistical measures and show their impact on the accuracy of the classifiers.

References

[1] Abdelhamid N., Ayesh A., and Hadi W., "Multi-Label Rules Algorithm Based Associative Classification," *Parallel Processing Letters*, vol. 24, no. 1, pp. 1-21, 2014.

[2] Abdelhamid N., Ayesh A., and Thabtah F., "Emerging Trends in Associative Classification Data Mining," *International Journal of Electronics and Electrical Engineering*, vol. 3, no. 1, pp. 50-53, 2015.

[3] Ajlouni M., Hadi W., and Alwedyan J., "Detecting Phishing Websites Using Associative Classification," *European Journal of Business and Management*, vol. 5, no. 15, pp. 36-40, 2013.

[4] Al-Fayoumi M., "Enhanced Associative Classification Based on Incremental Mining Algorithm," *International Journal of Computer Science Issues*, vol. 12, no. 1, pp. 124-130, 2015.

[5] Alazaidah R., Thabtah F., and Al-Radaideh Q., "A Multi-Label Classification Approach Based on Correlations Among Labels," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 2, pp. 52-59, 2015.

[6] Alwidian J., Hadi W., Salam M., and Mansour H., "Categorize Arabic Data Sets Using Multi-Class Classification Based on Association Rule Approach," *in Proceedings of the International Conference on Intelligent Semantic Web-Services and Applications*, Amman, pp. 18, 2011.

[7] Alwidian J., Hammo B., and Obeid N., "WCBA:

Weighted Classification Based on Association Rules Algorithm for Breast Cancer Disease," *Applied Soft Computing*, vol. 62, pp. 536-549, 2018.

[8] Alwidian J., Hammo B., and Obeid N., "Enhanced CBA algorithm Based on Apriori Optimization and Statistical Ranking Measure," *in Proceeding of the 28th International Business Information Management Association*, Seville, pp. 4291-4306, 2016.

[9] Alwidian J., Hammo B., and Obeid N., "FCBA: Fast Classification Based on Association Rules Algorithm," *International Journal of Computer Science and Network Security*, vol. 16, no. 12, pp. 117-127, 2016.

[10] Brooks J., "Anti-Phishing Best Practices: Keys to Aggressively and Effectively Protecting Your Organization from Phishing Attacks," *White Paper, Cyveillance*, 2006.

[11] Gupta S. and Singhal A., "Dynamic Classification Mining Techniques for Predicting Phishing URL," *in Proceeding of Soft Computing: Theories and Applications*, Singapore, pp. 537-546, 2018.

[12] Hadi W., "EMCAR: Expert Multi Class Based on Association Rule," *International Journal of Modern Education and Computer Science*, vol. 5, no. 3, pp. 33-41, 2013.

[13] Hadi W., Aburub F., and Alhawari S., "A New Fast Associative Classification Algorithm for Detecting Phishing Websites," *Applied Soft Computing*, vol. 48, pp. 729-734, 2016.

[14] Hadi W., Salam M., and Al-Widian J., "Performance of NB and SVM Classifiers in Islamic Arabic Data," *in Proceedings of the 1st International Conference on Intelligent Semantic Web-Services and Applications*, Amman, pp. 1-6, 2010.

[15] Hota H., Shrivas A., and Hota R., "An Ensemble Model for Detecting Phishing Attack with Proposed Remove-Replace Feature Selection Technique," *Procedia Computer Science*, vol. 132, pp. 900-907, 2018.

[16] Kulkarni M., Varma K., Patel S., Mer U., Parmar S., and Mahajan A., "A Study of Phishing Detection Using Associative Data Mining," *International Journal Of Scientific Research in Science, Engineering and Technology*, vol. 4, no. 5, pp. 419-423, 2018.

[17] Li W., Han J., and Pei J., "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules," *in Proceedings of IEEE International Conference on Data Mining*, San Jose, pp. 369-376 2001.

[18] Liu B., Hsu W., and Ma Y., "Integrating Classification and Association Rule Mining," *in Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, New York, pp. 80-86,1998.

[19] Ma B., Zhang H., Chen G., Zhao Y., and Baesens B., "Investigating Associative Classification for Software Fault Prediction: An Experimental Perspective," *International Journal of Software Engineering and Knowledge Engineering*, vol. 24, no. 1, pp. 61-90, 2014.

[20] Magazine F., Online Shopping Worldwide Ecommerce Statistics, Last visited, 2011.

[21] Parekh S., Parikh D., Kotak S., and Sankhe S., "A New Method for Detection of Phishing Websites: URL Detection," *in Proceeding of International Conference on Inventive Communication and Computational Technologies*, Coimbatore, pp. 949-952, 2018.

[22] Pereira R., Plastino A., Zadrozny B., and Merschmann L., "Categorizing Feature Selection Methods for Multi-Label Classification," *Artificial Intelligence Review*, vol. 49, no. 1, pp. 57-78, 2018.

[23] Rao C., Ramana A., and Sowmya B., "Detection of Phishing Websites Using Hybrid Model," *GPH-Journal of Computer Science and Engineering*, vol. 1, no. 1, pp. 15-22, 2018.

[24] Sankhyan R., Shetty A., Dhanopia L., Kaspale C., and Dantal P., "PDS-Phishing Detection Systems," *International Research Journal of Engineering and Technology*, vol. 5, no. 4, pp. 2429-2431, 2018.

[25] Sriramoju S., Ramesh G., and Srinivas B., "An Overview of Classification Rule and Association Rule Mining," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, no. 1, pp. 1962-1970, 2018.

[26] Tan P., Steinbach M., Kumar V., Taware S., Ghorpade C., Shah P., Lonkar N., and Bk M., "Phish Detect: Detection of Phishing Websites Based on Associative Classification (AC)," *International Journal of Advanced Research in Computer Science Engineering and Information Technology*, vol. 4, no. 3, pp. 384-395, 2015.

[27] Taware S., Ghorpade C., Shah P., Lonkar N., and Bk M., "Phish Detect: Detection of Phishing Websites Based on Associative Classification (AC)," *International Journal of Advanced Research in Computer Science Engineering and Information Technology*, vol. 4, no. 3, pp. 384-395, 2015.

[28] Wadhawan R., "Prediction of coronary heart disease using Apriori algorithm with data mining classification," *International Journal of Research in Science and Technology*, vol. 3, no. 1, pp. 1-15, 2018.

[29] Witten L., Frank E., Hall M., and Pal C., *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2005.

Mustafa Al-Fayoumi received a BSc degree in Computer Science from Yarmouk University, Irbid, Jordan, in 1988. He earned an MSc degree in Computer Science from the University of Jordan, Amman, Jordan, in 2003, and his PhD in Computer Science from the Faculty of Science and Technology at Anglia University, UK, in 2009. Currently, he is the Dean's Assistant for King Hussein School of computing sciences at Princess Sumaya University for Technology (PSUT), Jordan, His research interests include computer security, cryptography, identification and authentication, wireless and mobile networks security, e-application security, simulation and modelling, algorithm analyses and design, information retrieval, data mining and other related topics.

Jaber Alwidian holds a PhD in Computer Science (The University of Jordan). He received his B.Sc. degree in Computer Information System from the University of Philadelphia and M.Sc. degree in Information System from the Jordan University in 2005 and 2010, respectively. He has about seven years of work experience as a lecturer and one year as a big data scientist (INTRASOFT Middle East/big data department). His research interests are data mining, software engineering and image processing.

Mohammad Abusaif received a BSc degree in Computer Science from the University of Jordan, Amman, Jordan, in 1994. He is a senior level Manager with 24 years' experience working within the IT industry and software development, ERP, EMR and Big Data implementation of complex business systems in Jordan, United Arab Emirates, Kingdom of Saudi Arabia, Qatar, Oman, Lebanon and Yemen using various software and hardware platforms. Experienced at building, managing, motivating and leading multi-cultural teams, both local and globally distributed, while delivering complex projects covering various business sectors including: Government sector, Constructions, private sector, Healthcare, Retail and Distribution Business.