An Associative Classification Approach Using Data Mining for Predicting Phishy Email

BY

Bader Abdul Hafeez Al-rahamneh

SUPERVISOR

Dr.Fadi Thabtah

This Thesis was Submitted in Partial Fulfillment of the Requirements for the Master's Degree in Computer Science

Deanship of Academic Research and Graduates Studies

Philadelphia University

January 2013

جامعة فيلادلفيا

نموذج التفويض

انا بدر عبدالحفيظ احمد الرحامنه ،أفوض جامعة فيلادلفيا بتزويد نسخ من رسالتي للمكتبات أو المؤسسات أو الهيئات أو الاشخاص عند طلبها.

التوقيع:

التاريخ: 2013\01\02

Philadelphia University

Authorization Form

I am Bader AbdulHafeez Ahmed Al-Rahamneh, authorize Philadelphia University to supply copies of my Thesis to libraries or establishment or individuals upon request.

Signature:

Date: 02/01/2013

An Associative Classification Approach Using Data Mining for Predicting Phishy Email

BY

Bader AbdulHafeez Al-Rahamneh

SUPERVISOR

Dr.FadiThabtah

This Thesis was Submitted in Partial Fulfillment of the

Requirements for the Master's Degree in Computer Science

Deanship of Academic Research and Graduates Studies

Philadelphia University

January 2013

Successfully defended and approved on 02-JAN-2013

Examination Committee	Signature
Dr. Fadi Thabtah, Chairman. Academic Rank: Associate Professor.	----------
Dr. Hasan Alrefai, member. Academic Rank: Assistant Professor.	----------
Dr. Ibrahim Al-Oqily , External Member. Academic Rank: Assistant Professor. (Name of University: The Hashemite University).	----------

Dedication

For my family,

Who gave me endless love,

Support and encouragement

Throughout the course of this thesis.

Bader Al-Rahamneh

ACKNOWLEDGEMENT

No one could reach anything without the help of great Allah, so for what I have become today and for all of Allah graces he gave to me, I thank Allah almighty.

I'm heartily thankful to my supervisor, Dr.Fadi Thabtah, whose encouragement, guidance and support from the initial to the final level enabled me to develop and understanding the thesis.

It is my honor to express my thankfulness to my parents ,brothers and sisters for their support to accomplishing this thesis successfully.

Never forget to thank my wife Tamara,my daughters Maryam & Zena and my friend Shaima Alsatari or anyone that leads me towards my goal to achieve this work.

Bader Al-Rahamneh

List of Contents

List of Tables

List of Figures

List of Abbreviations

ACRONYM/SYNONYM	MEANING
AC	Associative Classification
ACCF	Associative Classification Based on Closed Frequent Itemsets
ARFF	Attribute Relation File Format
BART	Bayesian Additive Regression Trees
CACA	Class Based Associative Classification Approach
CARs	Class Associations Rules
CART	Classification and Regression Trees
CBA	Classification Based on Association Rule
CBART	Classification Bayesian Additive Regression Trees
CMAR	Classification Based on Multiple Class-Association Rules
DNS	Domain Name System
ECMC	Enhancing of the Evolving Clustering Method for Classification
LR	Logistic Regression
MCAR	Multi Class Associative Classification
NNet	Neural Network
ODBC	Oracle Database Container
PECM	Phishing Evolving Clustering Method
PINs	Personal identification numbers
RF	Random Forests
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
SCPE	Save Cyber from Phishy Email
SQL	Structured Query Language
SVM	Support Vector Machine
UCI	University of California Irvine

ABSTRACT

Email communication has come up as an effective trend of communication nowadays. People are sending and receiving many messages per day, communicating with each other and interchanging files and information. Phishing using email is a common electronic crime. It is one of social engineering techniques used to get advantage of human unawareness. It allows abusive people to utilize the weaknesses in web security technologies to get confidential information, such as usernames, passwords, financial account credentials and credit card details. This thesis handles phishy email problem usingclassification based on association rule mining, which is a common data mining approach that merges association rule discovery in the learning step within classification techniques. Experimental studies manifest that this learning approach produces clear simple rules by discovering all possible correlations among all attributes which is output "If-Then" classifiers that are understandable by end-user.However, this approach generatesmany rules that may redundant so this thesisproposes a new rule ranking method, rule pruning procedures in classifier building phase to remove unnecessary rules without impacting accuracy rate and new predicting procedure to obtain high accuracy rate. The result is an algorithm that discovers rules from dataset to improve upon previous works. The proposed algorithm comprises the following characteristics:

1. A pruning procedure that minimizes the size of the classifiers, to come up with controllable number of rules.
2. A prediction procedure that employees multiple rules for assigning the class for test data rather than using a single rule prediction as the majority of current techniques hence the accuracy rate improved on the classifiers.

The proposed algorithm has been applied on the email problem and compared withDecision Trees, NaiveBayes and Rule Induction against real dataset.The result showed superiority of our associative classification algorithm especially in outperforming the rest of algorithms by accuracy rate. In particular, the algorithm produced higher accurate rate classifier. The email dataset used in the experiments consists of one thousand instances and contains eight features extracted manually.

Chapter One

Introduction

1.1 Introduction

Email communication has come up as an effective trend of communication nowadays. People are sending and receiving many messages per day, communicating with other people, or interchanging files and information. Phishing charges using email are the most common of electronic crime (Activity, 2012; Khonji et al.,2011).It is one of social engineering techniques used to get advantage of human unawareness. It allows abusive people to utilize the weaknesses in web security technologies which try to get confidential and private information, such as usernames, passwords, financial account credentials and credit card details, by veiling as a proper object in an email.

Lightheaded cyber user may be easily deceived by this kind of scam. Victims of phishing email may lose their bank account details, password, credit card number, or other private information to the phishing email senders. In addition phishing is considered as spam; while it is being differs from spam (Irani, et al., 2008). Indeed spam almost seeks to sell a product or service, while a phishing message try to look like it is a form of legitimate organization. Straightforward, approaches that are handled the spam messages cannot be used to phishing messages (SonicWall, 2008; Irani et al., 2008).

According to the Anti-Phishing Working Group (APWG) reports for the first quarter on 2012 (Activity, 2012) shows a terrible number of phishy emails attack, there were on average 28,481 unique reports of phishy emails (campaigns) received by APWG from web users. The email campaign is a unique email is sent out to multiple users mislead them accessing a specific phishy website. Indeed, financial services found to be the most-targeted industry sector in the first quarter of 2012. Moreover, Payment Services eclipsed retail/services have the second-highest industry sector for targeted attacks (Activity, 2012). FraudAction Research Labs divulge that, phishing attacks on the first half of 2012 have been increased compared by the same period of 2011 (Kovacs, 2012).

1.2 Motivation

Phishy email is not a relaxed issue to handle and understand or even to analyze, since it combines technical and social aspect where no silver bullet exists to solve it in straightforward manner. That is why we attempt to quantify and qualify the phishy email features in order to understand the protective measures to prevent or mitigate the risks and threats that come from phishy email especially the creation of the “trust crises” which severely affects all online transactions.

Phishy emails cause suffering from different losses for users either in personal or as banks, online business or social network sites.It usessocial engineering approach, which can be defined according to (Damodaram et al., 2012) as “ when phishers find exploiting human nature is easier than exploiting weakness in software using human emotional to acting fraudulently by dealing with users through performing actions or spread confidential information”. Moreover, this kind of technique always uses for the purpose of information collection, scam, or computer system access. Figure 1.1 shows an example for malware resulted after installing it on the machine and how phishers can collect users' privet information such as credit card numbers while they lure the user by selling him anti-virus software called "Windows Health Keeper". However, social engineering not only the way phishers follow to lure their victims but different technical techniques are used such as put a malware in images or in attached file. Payment services eclipsed retail/services have the second-highest industry sector in targeted attacks. In fact, $687 million were what the worldwide paid for this kind of electronic crime (Kingdom, 2012). Also hijacked emails is one of the very dangers scams (Activity, 2012).

Providing a resilient effective and safe web environment by detecting phishy emails in order to help users from being deceived or hacked in terms of their personal information is crucial. Indeed, classify such kind of problem offers many benefit on personal and commercial level. To the best of our knowledge, phishy email problem has not been handled by classification based on association (AC) data mining to assess the type of emails. Moreover, features selection determines the quality and effectiveness of the classification system (Ma, et al., 2009) and thus choosing the right features is also a concern of this thesis.

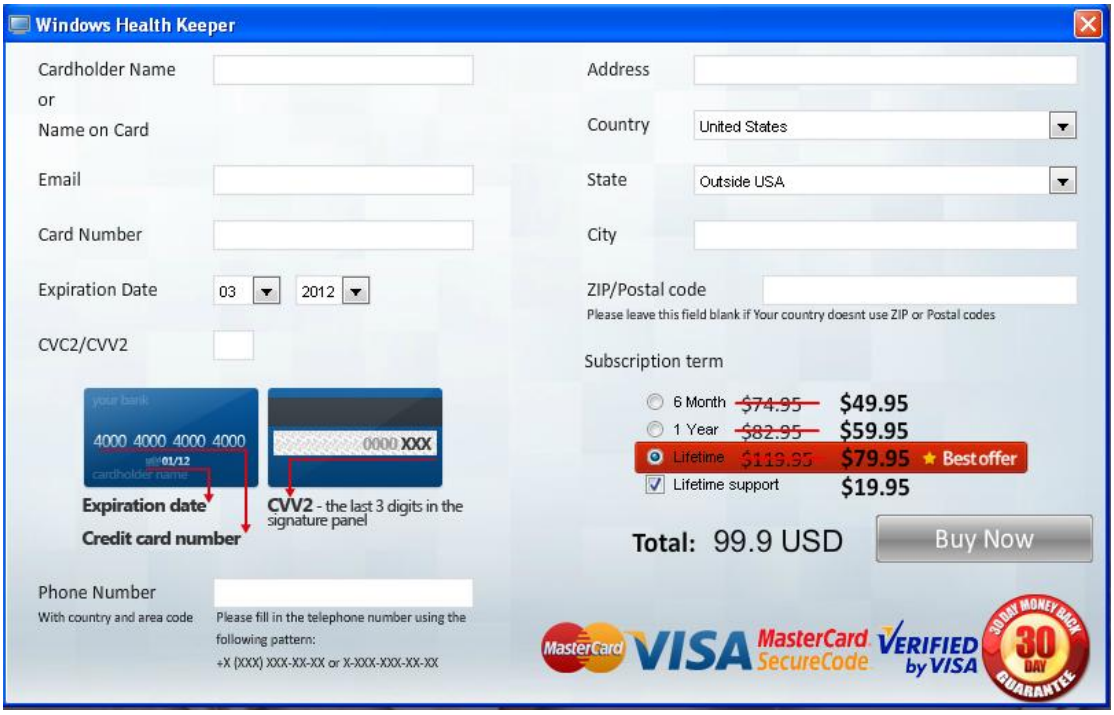

Figure 1.1 Malware Result(Spywareremove, 2013)

In particular, we are intending to focus on the header based features such as, subject reply, subject verify, etc, and the content based features such as HTML, long URL addresses, etc, and then select those which offer higher classifyingquality against emails data. Using AC brings us a correlation of different attributes in simple knowledge format yet effective knowledge. AC always finds all relationships among attributes values (Bjorn et al., 2011), allowing it to be applied in different classification problems(Baralis et al., 2008).

1.3 Phishing History

In 1996 the word “Phishing” was first mentioned as a combination of ”password” and ”fishing” on the internet in the hacker newsgroup (Martino et al., 2010). The idea is that bait is thrown out with the hopes that the user will grab it and bite into it just like the fish. In most cases, bait is either an e-mail or an instant messaging site, which takes the user to hostile phishing websites (Razvan et al., 2010). The fishermen usually camouflage to be known bank, tradesman on line, corporations of credit card and so on.

Phishing and spam email differ in their goals and targets. They use different features to achieve their aspects. Anti-spam software is much for detecting and handling non-targeted spam email. On the other hand, phishy email is harder to detect

(Guofei et al., 2007) since it combines between social engineering and technical techniques. It helps to limit what personal information you share online, such as on social networks. Therefore, anti-spam software acts with low accuracy when it is used to predict phishing emails(SonicWall, 2008).

1.4 Data Mining

Mining the warehoused data to find out the interesting patterns and the associations in the data is important(Karthikeyan et al., 2012). Data mining is the science of extracting meaningful information from these large data sets (Saad et al., 2011). Data mining and knowledge discovery techniques have been employed to different areas including market analysis, industrial retail, decision support and financial analysis (Toolan et al., 2010). Association rule mining and classification rule mining are two important data mining techniques. Classification and association rule discovery are akin unless that classification exercise prediction of one attribute, i.e., the class, on the other hand, association rule discovery can describe any attribute in the data set.

There are several major data mining techniquesthat have been developed and used including association rule, classification, clustering and regression (Gupta et al., 2012). Below we briefly examine those data mining techniques with example to have a good overview of them.

1.4.1 Association Rule

Association rule is one of the known data mining techniques (Gupta et al., 2012). This approach offers patterns based on a relationship between different items (Chen et al., 2005). Market basket analysis is an obvious example for the association approach when it is used to identify what products frequently purchase together by customers. As a result, businesses can have corresponding marketing campaign to sell more products and to make more profit. Also it helps them in decision making processes such as shelving.

1.4.2 Classification

Essentially it is used to categorize items in a set of data into a predefined set of classes (Thabtah et al., 2005). Mathematical techniques such as decision trees, linear programming, neural network and statistics are used with this approach. We can apply classification in application that “given all records of employees who left the

company, predict which current employees are probably would like to leave in rare future.” In this case, we divide the employee’s records into two groups, “leaving” and “staying”.

1.4.3 Clustering

It is similar to classification but in an unsupervised way. Clustering defines the groups and put objects in them, while in classification objects are assigned into predefined classes that makes significant group of objects share similar characteristics(Al-Momani et al., 2011). In a library for example, books have a wide range of topics available. The challenge is how to gather those books in a way that readers can take several books in a specific topic without extensive search or effort. Clustering introduces some kind of similarities in one cluster or one shelf and arranges it with a meaningful class. Consequently, readers just go to that shelf instead of looking in the whole library.

1.4.4 Regression

This technique has two styles, the simplest is linear regression, utilizes the formula of a straight line (y = mx + b) and specify the proper values for m and b to forecast the value of y depending on x. on the other hand, the advanced, like multiple regression, admit the use of more than one input and admit for the fitting of more complex models, such as a quadratic equation (Kenkel et al., 2011).

1.5 Problem Statement

Phishing attacks are designed to steal confidential and private details from users. As a result the cyber criminals can assume controlling the victim’s in social network, email accounts, and online bank accounts. This can be done because criminals use login details, to access multiple private accounts and manipulate them for their own good (Salama et al.,2012).Moreover, this problem creates a “*trust crises*” which severely affects all the online transaction (Data et al., 2012). Phishers always try to render emails looking like legitimate when all they actually need is the user personal information which causes losing of large amount of money in different business.

Some users may are aware of phishing but this problem could lure easily users with high educating or even technical skills. This is actually one of the newest scams that computer hackers use to gain access to another person's confidential and private

information. Phishing is one of the hottest and fiercer topics in the field of identity theft (Asanka et al., 2012).

This research goal is to investigate the potential use of automated data mining techniques in detecting the problem of phishyemail. Particularly, we aim to develop a based rule data mining model that will be applied to predict (classify) the type of email as accurately as possible. This is a binary classification problem which let us classify the email in two specific labels “phishy” or “legitimate”.

Moreover, AC derives massive number of rules in the form “if-then” since it basically discovers every single correlation between the attribute values and the class attribute in the training dataset(Bjorn et al., 2011).We aim to develop a rule ranking method, rule pruning procedures in classifier building phase to remove unnecessary rules without impacting accuracy rate. In addition, enhancing the process of forecasting the type of email by using more than one rule is done in this thesis. The classification process will be based on the different features such as HTML, long URLs, etc, collected from the input email. In the result, cutting down the classifier size after implementing the new procedures the end-user ends up with less sized classifier than these in other techniques in which he can easily understand and maintain.

1.6 Research Questions

The following are the research questions to be answered in this work:

- Is AC data mining an accurate approach for detecting email type?
- Does reducing number of rules produced by our model impact the accuracy of the model?
- What are the significant set of features that can detect the type of email effectively?
- Is group of rules prediction appropriate for accuratly detecting the type of an email.

1.7 Thesis Contributions

The thesis goalis to build phishing detection model that uses rule based data mining methods to find out whether phishing activity is taking place on an email. The resulting implementation has to be effective and practical, and has to produce accurate identification for example, avoiding false-positive and false-negative.

- This implementation is based on email's features which have been selected after analyzing them in Section (3.3). The assessmentaims to obtain the significant features set that the model can train on to derive classification with a high classification with high accuracy rate.
- The model is built with a rule ranking offering superior, high confidence and high support rules in the classifier for the purpose of applying them later in the prediction phase. Superior rule has the maximum number of attributes which is a specific rule that are more accurate in predicting test instance especially because they cover smaller number of training cases.
- Also a new rule pruning applies partial matching as new criteria if full matching of the candidate rule body and the training set is not met. This technique offers a classifier contains less number of rules because a rule now has more training instance coverage.
- Moreover, a new forecasting step is presented in order to classify unseen instance class by depending on mathematical formula, Chapter 3 handles details for the new procedures and Figure 1.2 shows the general process of the proposed model.

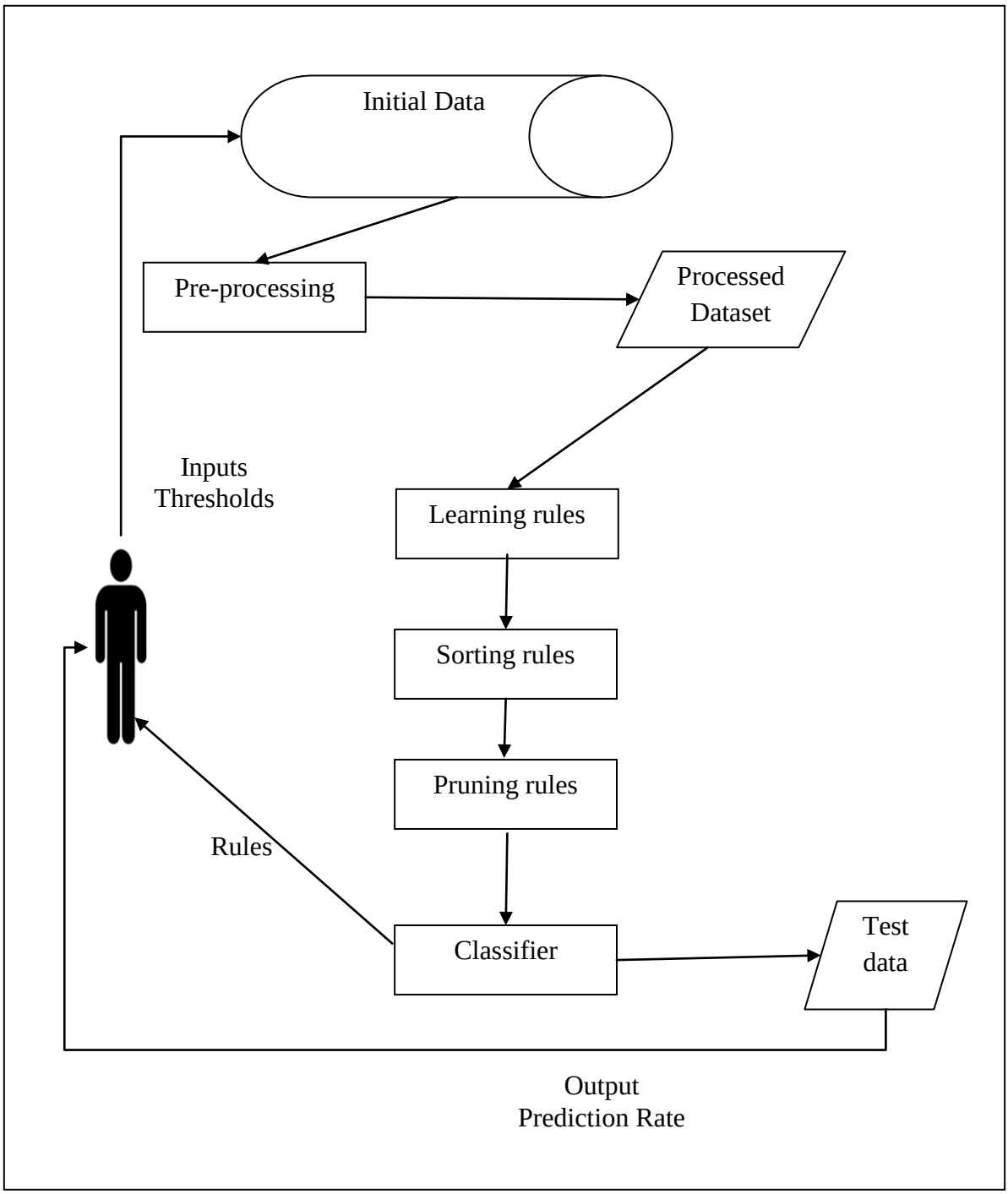

Figure 1.2: General step in the proposed model

1.8 Thesis Outline

The thesis is structured in five chapters. Chapter two describes currently developed techniques that are used in data mining techniques to eliminate or reduce the phishy email problem. Also, it presents the related works that are used different techniques with focusing on features that are used by the researchers. Common AC algorithm was also concern of this chapter by focusing in some problem upon AC

approach. Chapter three handles the analysis for different related features that has been investigated by frequency analysis as well as the proposed model as a new method to enhance a combined approached called AC to deal with phishy email problem was presented in this chapter. Chapter four handles the experiments and results after compared with other methods such as NaiveBayes, J48 and Prism regarding to accuracy rate, average Precision and Recall and number of rules which are produced. Lastly, we conclude this work and shed the light on some further research directions in Chapter five.

Chapter Two

Literature Review

2.1 Introduction

Several ways are considered to pounce upon phishing. These rotate from communication-oriented approaches like authentication protocols over blacklisting to content-based filtering technique (Paaß et al., 2009). The first two techniques are currently not that much implemented or exhibit deficits. Blacklists, Whitlists or both of them are not that qualified when used in different filters approaches, while continuously a fresh phishy scam is engendered. In fact, these filters are dropped in scalable problems. Therefore content-based phishing filters are requisite and ample used. Soresearchers focus on machine learning and data mining techniques to carry this problem based on the email contents in the header and on the body of an email.

If phishing could be completely eliminated using these methods, there would be no need for other protection strategies. However, existing tools are unable to detect phishy emails and phishy websites with 100% accuracy (Kumaraguru, et al., 2010). For example, in 2007, study showed that even the best anti-phishing toolbars miss over 20% of phishing websites and study in 2009 found that most anti-phishing tools did not start blocking phishing sites until several hours or days after phishing emails had been sent luring users to those sites (Parmar, 2012).

This chapter is covered different techniques that are applied in order to detect and predict phishy emails under two types, traditional methods and automated methods techniques which are handled in Section 2.2. Compression among these different techniques is presented in Section 2.3. Section 2.4 intercourse with phishy emails related work and their features. Associative classification is handled in Section 2.5and itsrelated definitions are presented in Section 2.6. The common AC algorithms are handled in Section 2.7. Finally, we summarized the chapter in Section 2.8.

2.2 Techniques to Handle Phishy Email

Researchers have developed filters depended on different techniques to eliminate the phishy email problem based on traditional techniques such as network level protection and authentication protection as well as on modern techniques using machine learning and data mining approaches.

2.2.1 Traditional Methods

Classifiers in this type could be categorized to two natures of classifiers. Firstly, network level protection such as Blacklist filters and Whitlist filters. This type is remedied through blocking ambit of IP addresses or collection of domains from passing the network. Moreover, Pattern Matching filters and Based Rule filters are handled by fixed rules which are required continuously updated manually. On the other hand, Authentication protection has two levels, user level and domain level is the second category. The latter is between email servers while the former obliges users to have an authentication before they send their messages (Ramanathan et al., 2012) such as Email verification filters and Password Filters.

❖ Black List Filter

Black list is a network level protection that used as a filter to classify email as phishy or legitimate. This technique is examined the sender's address, IP address or DNS address by extracting these data from the email header with predefined list. If any one of these data has match what in the list it is rejected (Sheng et al, 2009), then this email is classified as phishy so it is not received by their recipient (Paaß et al, 2008). Internet Server Providers (ISP) is the one who responsible for applying this procedure.

❖ White List Filter

This type of filters is connected with lists having static IP addresses for legitimate domains at network level protection. It is compared the email address with the IPs addresses that are found in the white list (Cao et al, 2008). If the matching step handled a positive result then the email bypasses the filter and goes to the receiver's inbox. This white list is filled by emails received from legitimate companies or people who agree their addresses or IP addresses to be included in the list so this way the sender's identity always is known. This filter is categorized as legitimate emails classifier because it is based only on the legitimate address. However, every email is not considered in the white list is classified as a phishy email leading to suffer or incur loss significant emails. (Paaßet al, 2008).

❖ Pattern Matching

Filters are based on some concepts that classify the email as either phishy or legitimate at network level protection, by searching for specific words, text strings and character sets in the email content, sender and subject line. But there is huge number of false positives in the result since many emails contain banned words or text strings (Shalendra Chhabra, 2005).

❖ Email Verifications

This technique is user level authentication which has two agents, sender and receiver. The acceptance of the message is verified when the receiver sends a notification message back to the sender to identify the message (Adida et al, 2006), which means the message is legitimate so it is passed to the inbox. Otherwise the message is classified as phishy email and does not pass. The advantage of this technique is the ability to filter almost 100% of the phishy emails. On the other hand, two limits based on this technique appear; first it is time consuming because the receiver needs to respond. Secondly, if this challenge is not recognized the email will be lost, and the verification email also generates more traffic over the network.

❖ Password Filter

Passwords are embedded to receive any email in the subject line, the email address, the header field, or in any part in the email at user level authentication. When the filter finds the password, the email surpasses through the filter. The email is rejected due to the fact that the password is not recognized. when first time using this approach users need to initiate a conversation with each other because the email does not yet contain a password, so asking new user to allow their password is needed to pass through the filter causing time consuming or losing legitimate emails if the password is not given (Ramanathan et al., 2012).

2.2.2 Automated Methods

Automated classifiers are a server side filters and classifiers based on machine learning and data mining approaches. Extracting different features from the email header and body is the current method followed to process the classifier depending on them allowing distinguish between emails if phishy or legitimate. Two most applied approaches types of classifiers, statistical such as Bayesian and Support Vector

Machine (SVM). Second type is multi-layer classifiers such as Decision Trees (DT) and neural network.

2.2.2.1 Statistical Based Methods

Classification on the basis of training set of data containing instances which class attribute is known to classify unseen case.

❖ Bayesian Filter

This filter belongs to statistical text classification systems it is based on a Thomas Bayes' theory that became a popular formula in 1950. This theory was used in many fields of science, and it is useful in the computer field. This theory depends on the previous event to prove the conditions and give the optimal solution to solve the problem. It is a relationship among conditional and minor probabilities. It can be perceived as a method of combining or merging information (Shih et al., 2006). The application of Bayesian theory for phishing email detection is useful because it can be used to identify certain features in the email messages, how often they occur, and the probability that each message is a phishy email. By storing this information in a database this data can then be used to predict the probability of unseen instance.

❖ Support Vector Machine Filter (SVM)

SVM is a statistical technique to distinguish two different categories of data using particular rules, also deal with Quadratic equations. It is now adaptive in many fields such as medical diagnoses, text categorization, image classification, bio sequences analysis, etc. SVM builds a hyper plane to arrange data into two categories, depending on maximization the distance of the margin base on kernel functions to find the ideal solutions so it could be used as a binary classification (Abu-nimeh et al., 2007)(Burges, 1997). Indeed, it extracts the features and stores them in a vector, then uses them to classify the data depending on the problem. SVM is considered as one of the popular statistical techniques in classification while it has no past information about the problem. However, a limitation with SVM is appeared when the size of data is enormous causing memory consuming.

2.2.2.2 Multi-layer Methods

This structure helps to improving error-rate performance in different classification tasks. Multi-layer structures has a parallel development, and results are presented to characterize the performance of such structures in a practical character

recognition environment for a range of configurations and, in particular, for hierarchies of increasing order (Ramanathan et al., 2012).

❖ Decision Trees Filter (DT)

DT is based on If-Then rules. It is a graphical classifier that contains many nodes connected with each other by arrows called edges. Each tree starts from the first node called the root. This node is the base of the decision tree. It has one classifier stage or many classifier stages. Every decision trees has a leaf node to terminate the tree. Moreover, the internal nodes are the nodes between the root and the terminator nodes (Safavian et al., 1991).

Each node in the tree contains a decision rule, class, and feature. There are many classifier algorithms proposed. ID3 is a classifier algorithm that was proposed which calculate information entropy as heuristic function to evaluate target, then in (1992) the C4.5 algorithm was proposed, and this algorithm can be considered as an extension of ID3 algorithm (Olaru et al., 2003). Using these algorithms will generate the decision tree from the beginning and it will contain sub trees, every nodes has parent except the root, and each one of them has child node except the leaf nodes, the leaf nodes is the solution of the problem.

❖ Neural network Classifier

A neural network contains of a number of similar linked neurons. These neurons are interacted with each other to pass the information. This interacted has weights to prove the delivery between neurons which are not usable in single mode. On the other hand, they can deal with difficult problem if they interacted. Neurons weights are changed by interconnections when the network is processed. Figure 2.1shows a neural network that including, one input layer, one hidden layer that nonlinear which reflects the strength of neural network, and one output layer. The nonlinearity is important in the neural network to learn complicated mappings (Abu-nimeh et al., 2007).

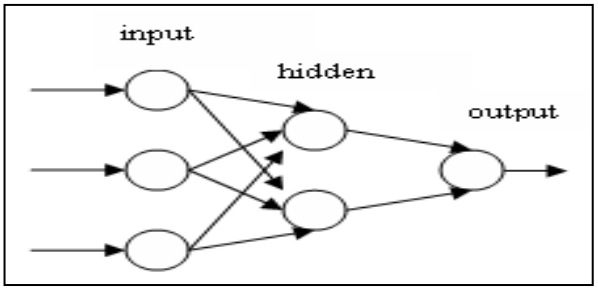

Figure 2.1: Neural Network (Abu-nimeh et al., 2007)

2.3 ContrastingAmong the Previous Techniques

All the previous techniques have advantages and disadvantages are shown in two tables, Table 2.1 for the traditional methods and Table 2.2 for the automated methods.

Table 2.1 Comparison between traditional methods for the phishy email filter techniques

Technique name	Advantage	Disadvantage
Blacklist filter	Blocks determined phishy email	Dose not block new phishy email
White list filter	Blocks email form unknown senders	Block new legitimate email
Pattern matching	Easily blocks emails based on predefined pattern	Has a huge number of false positives in the result
Email verifications	Ability to filter almost 100%	Time consuming, email will be lost when no response and generates more traffic on network
Password filter	Allow only the email that has password passing through	Block legitimate email that does not have a password yet

Table 2.2 Comparison between automated methods for the phishy email filter techniques

Technique name	Advantage	Disadvantage
Bayesian filter	Calculate the probability of the message phishy or not. Self-learning technique	Does not deal with HTML or image mail while it belongs to statistical text classificationsystem
SVM filter	Has no past knowledge about the problem. Self-learning technique	Needs time and memory when size of data is large
Decision tree filter	Very simple to interpret and to understand by humans.	Creates redundant data
Neural network classifier	Achieved complex computations	Deal with nonlinearity to learn complex mappings

2.4 Related Work and Features

Accuracy rate and false positive rate are the main measures that researchers employed to be the significant concrete of filter's utility. On the other hand, features count and computational process to predict phishy emails are unusually considered (Olivo et al., 2011). Some existing techniques are emerged to haphazard choosing of the features that are used in classification (Toolan et al., 2010;Islam et al., 2009). Phishing filters are liable for applying features that are directly belonged to phishy emails than those used by general purpose such as in spam filters. We have presented

most technique that have been adopted by researchers in Section 2.4.Discarding important features in training deteriorates the learning performance, and on the other hand including redundant ones performs to over-fitting(Ma et al., 2009).

Moreover, filters based on automated methods are consisted for the following steps below (Shalendra ,2005) .

1. Corpus: phishy and legitimate emails are collected and assented.
2. Feature Generation: Features (attributes) are shed light upon each class phishy or legitimate are generated. So anything in the email could be applied as a feature.
3. Classifier Training: using a machine learning approaches to train the filters based on emails features discovered above.
4. Threshold: the classifier is performed to extract the features to classify unseen instance, phish or legitimate.

PILFER is the first Machine Learning based email classifier to expose phishy email based on support vector machine techniques, which showed promising results that acquired low false positives while prevent using blacklists (Khonji et al., 2011). It is a new method for exposing malicious emails by combine features specifically designed to highlight the delusive methods used to scatter users (Amelia et al., 2011). About 92% of phishy emails are correctly classified, while 0.1% is the rate of false positive. There are 860 phishy and 6950 legitimate emails are used to train the classifier. The fineness of PILFER on this dataset is consequence better than that of SpamAssassin, which is one of extensively used spam filter. As a result it is possible to expose phishy emails through a specialized filter, applying features that are more directly suitable to phishy emails than those used by general purpose such as in spam filters. Therefore, different 10 features were used in PILFER which are capableto achieve good accuracy rate, Table 2.3 shows PILFER’s features.

Table 2.3 PILFER features

Feature Id	Feature Name	Feature Description
1	IP-based URLs	Check if links have an IP-address
2	Age of linked-to domain names	Check the domain life
3	Nonmatching URLs	Check if different host than the link in the text
4	“Here” links to non-modal domain	domain most frequently linked to the “modal domain”
5	HTML emails	Check MIME type of text/html
6	Number of links	Check number of links in the html part(s)
7	Number of domains	Count the number of distinct domains
8	Number of dots	Maximum number of dots (‘.’)
9	Contains javascript	Check “javascript” string
10	Untrained SpamAssassin Output	Check if SpamAssassin labels an email as spam

Classification approach is one of the most used for content based filters. This approach is built its classifier over different features. In (Toolan et al., 2010), the most used features are categorized into five different groups, Table 2.4 handles the first group: Body-based features , Table 2.5 showsthe second group: Subject-based features, Table 2.6 handles the third group: Sender-based features which is proposed as the first time by (Toolan et al., 2010). Table 2.7 has the Script-based group. Table 2.8 and Table 2.9 handle URL-based features.

Table 2.4: Body-based features(Toolan et al., 2010)

Feature Id	Feature name	Feature description
1	Body html	Represents the presence of html in the email body
2	Body forms	Represents the presence of forms in html email bodies
3	Body no words	Measures the total number of words occuring in the email
4	Body no characters	Measures the total number of characters occuring in the email body
5	Body no distinct words	Measures the total number of distinct words occuring in the body
6	Body richness	Body no words /body no characters
7	Body no function words	Measures the total number of occurrences of function words in the email body such as account; access; bank; credit; click; identity
8	Body suspension	Represents the presence of the word suspension
9	Body verify your account	Represents the presence of the phrase verify your account

Table 2.5: Subject-based features(Toolan et al., 2010)

Feature Id	Feature name	Feature description
1	Subj reply	Records if the email is a reply to a previous email from the sender
2	Subj forward	Records if the email is forwarded from another account to the recipient
3	Subjnowords	Records the total number of words in the subject
4	Subjnocharacters	Records the total number of characters in the email's subject line
5	Subj richness	Subjnowords/subjnocharacters
6	Subj verify	Describes if the email's subject line contains the word verify
7	Subj debit	Describes if the email's subject line contains the word debit
8	Subj bank	Describes if the email's subject line contains the word bank

Table 2.6: Sender-based features(Toolan et al., 2010)

Feature Id	Feature Name	Feature Description
1	Send noWords	Represents the total number of words in the send field
2	Send noCharacters	Represents the total number of characters in the sender field
3	Send diffSenderReplyTo	Shows if there is a difference between the sender's domain and the reply-to domain
4	Send nonModalSenderDomain	Shows if the sender's domain is different from the email's modal domain

Table 2.7: Script-based features

Feature Id	Feature Name	Feature Description
1	Script scripts	Represents the presence of scripts in the email body
2	Script javaScript	Represents the presence of javascript in the email body
3	Script statusChange	True if the script attempts to overwrite the status Bar in the email client
4	Script popups	True if the email contains pop-up window code
5	Script noOnClickEvents	Feature counts the number of onClick events in the email
6	Script nonModalJsLoads	Represents the presence of external javascript forms that come from domains other than the modal domain

Table 2.8: URL-based features(Toolan et al., 2010)

Feature Id	Feature Name	Feature Description
1	URL IP address	Represents the use of IP addresses rather than a qualified domain name
2	URL no IP addresses	measures the number of links in an email that contain IP addresses rather than fully qualified domain names
3	URL at Symbol	represents the presence of links that contain an @ symbol
4	URL no links	measures the number of links in the email body
5	URL no Int links	Describes the number of links whose target is internal to the email body
6	URL no Ext Links	Describes the number of links whose target is outside the email body
7	URL no Img links	Measures the number of links where the user needs to click on an image in the email body
8	URL no Domains	Measures the total number of domains in all URLs in the email.
9	URL max No periods	measures the number of periods in the link with the highest number of periods

Table 2.9: URL-based features(Toolan et al., 2010)

Feature Id	Feature Name	Feature Description
1	URL link text	True if the human-readable link text contains one or more of the following terms: click; here; login; or update
2	URL non modal here links	Captures here links that Link to a domain other than the modal domain
3	URL ports	Indicates whether a URL accesses ports other than 80
4	URL no ports	Represents the the number of links in the email that contain port information in the address

Variously from predicting spam email number of research upon phishy emails are fewer (Thareja et al., 2011). However, a comparison among different machine learning techniques was presented and handling the accuracy rate including, Logistic Regression (LR), Classification and Regression Trees (CART), Bayesian Additive Regression Trees (BART), Support Vector Machines (SVM), Random Forests (RF) and Neural Networks (NNet) to give email type for unseen instance (phishy or legitimate) (Abu-nimeh et al., 2007). 2889 phishy and legitimate emails were tested which reflect the lateleaning in phishing. Therefore, the classifiers were trained based on 43 features to check them. Authors have used different evaluation metrics but let us shed the light on false positives and false negatives. Practically, false positives in this kind of problem is considered much costly than false negatives while users never want their legitimate email, which might be important, to be eliminated or wrongly classified. As a result, LR is obtained the bestclassifier than others that comes true when it is gained the lowest false positive rate. RF has the lowest false negative rate of 11.12%, followed by CART 12.93%, then LR got 17.04%, SVM rate was 17.26%, while BART 18.92%. Lastly, NNet has the highest false negatives rate which is 21.72%.

(Olivo et al., 2011) proposed that minimum number of appropriate features offers reliability, good performance and flexibility as a technique when predicting the phishy email. Authors appointed the Support Vector Machine (SVM) algorithm to deal with the proposed technique, which was firstly designed to combat binary classification problems. Identifying best features that establish the phishy emails was

the target authors focused on. Therefore, the techniques of the phishers were studied by searching for the minimum number of features that clearly illustrate the properties of phishy emails attack. As a result, literature is accorded identified the unanimous features, which are used by the majority of researchers with acceptable interpreter upon the phishing overview. This work was handled by testing the phishing dataset that could identify phishy emails if claim at least one of the eleven resulting features.

The instability structures and variability of email attacks cause current email filtering solutions useless. Indeed, the need for new techniques to rigid the protection of users' security and privacy becomes an essential. Therefore, (Abu-nimeh et al, 2009) provided a new version of Bayesian Additive Regression Trees (BART) and applied it to phishing detection. 1409 phishy emails and 5152 legitimate emails, where 71 features were itemized for the dataset. Analyzing both textual and structural on features was the responsible phase for generating the first 60 and the last 10 features in the dataset respectively. Moreover, six classifiers were applied to compare their efficiency according to phishy emails problem, the error rate, false positive (FP), and false negative (FN) rates measures were used with, Logistic Regression (LR), Classification and Regression Trees (CART), Bayesian Additive Regression Trees (BART), Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NNet). As a result, CBART reaped lowest FP rate of 2.98%, followed by RF with 4.24%, SVM with 5.43%, NNet with 6.16%, BART with 6.18%, LR with 7.29%, and CART with 11.55%. The minimum FN rate is reaped by CBART with 11.14%, followed by RF with 13.20%, SVM with 13.77%, NNet 14.32%, BART 16.48%, LR 18.38%, and CART 22.10%. So predicting phishy emails using the enhanced Bayesian Additive Regression Trees (CBART) algorithm is verified according to FP and FN.

profiling phishing vim was discussed as a novel approach through analyzing phishy emails in (Yearwood et al., 2010). Profiling is taken place in different groups either in individual or particular of phishers to sever if phishing found or not. Authors were innovated that profiling problem as a multi-label classification problem depending on hyperlinks in the phishy emails as a feature as well as structural properties of emails connect with Whois (DNS) information on hyperlinks as profile classes. AdaBoost and SVM algorithms were used to produce multi-label class predictions on three different datasets generated over hyperlink data in the emails.

The former provided very high classification accuracy rate since more accurate profiling was achieved. In (Ma et al., 2009) another approach for detecting phishy emails is used, a hybrid features are elected by information gain calculation. The election brought out three categories of features, Content, Orthographic and Derived handling seven features. Four components, Feature Generator, Machine Learning Method Selection, Inductor and Feature are included in the model for predicting phishy emails. Depending on five machine learning algorithm, Decision tree, Random forest, Multi-layer perceptron, NaiveBayers and support vector machine (SVM). As a result, the decision tree generated the highest accuracy rate and constructed a good classifier.

Enhancing of the Evolving Clustering Method for Classification (ECMC) is presented as a novel concept that brought out a new model named the Phishing Evolving Clustering Method (PECM) in (Al-Momani et al., 2011). PECM is processed between of reticulate two sets of features upon phishy emails, that reflect clustering-based learning model that enhance Evolving Clustering Method to distinguish between two types of emails phishy or legitimate in online form. PECM is demonstrated the tendency to category email by let the level of false positive and false negative in lowest rates while let the level of accuracy rate goes up to 99.7%. The proposed model include preprocessing, email object similarity and application of the clustering technique Phishing Evolving Clustering Method (PECM).

Table 2.10 summarize the popular phishing detection tools such as CloudMark, Netcraft, FirePhish, eBay Account Guard and IE Phishing Filter (Ramanathan et al., 2012). The authors pointed out the main disadvantages of the popular tools that are used. Indeed, we deal with all of them in this thesis to help users from be a victim of scam and identity theft.

Table 2.10 Phishing detections tools (Ramanathan et al., 2012).

Tool	Type	Description	Advantages	Disadvantages
Snort	Network level	Heuristic tool	Good at detecting level attacks	Rules require manual adjustments. Does not look at content
Spam Assassin	Server Side Filter	Heuristic engine uses specific features	Good at detecting email header spoofing	High false positives
PILFER	Server Side Filter	Utilize 10 features	Better performance than spam assassin	Did not use content from body of the email. Used with short lived phish domains.
Spoof Guard	Client Side Tool	Plug-in to a browser	Warns user if link points to phishing site.	Users do not pay attention to warnings. Not all email clients are browser based.
Calling ID, Cloud Mark, Netcraft, and Fire Phish	Client Side Tool	Utilizes blacklist of domains	Good for domains that employ domain level authentication	Phish domains are short lived. Does not look at email content.
eBay Account Guard	Client Side Tool	Utilizes blacklist of eBay URLs	Protects eBay users.	Specific website tool.
IE Phishing Filter	Client Side Tool	Records specific user website visiting patterns.	Adapts to user website visit pattern.	Works only on internet explorer.
Catching Phish	Client Side Tool	Detects fake website based on rendered images	Browser independent. Good results on small data sets.	Processing time is high. Susceptible to screen resolution

Tables 2.11 and 2.12 provide a list of features taken from (Toolan et al., 2010).This thesis focus on the features that are used only internally into the emails themselves by classify them to two groups, email header in first table and email body group in the second table. Many authors have used external features outside the

emails such as spam assassin scores, domain registry information, or search engine information (Toolan et al., 2010). Consequently, three major sides were beyond our depending on the email itself listed below.

1. In fact the email message is the only part of information that is liable for involved all people in the phishing detection task.
2. Instability of the external data, for instance DNS information or search results is modified with the time passing.
3. Blacklist and Whitelist methods are need too much work on the part of individuals and organizations and it is faced a scalability problem so that they are not trusty to invoke automated phishing detection system.

Table 2.11 Email header group

Id	Feature Name	Feature Description
1	Subj Reply	Binary feature records if the email is a reply to a previous email from the sender
2	Subj Forward	Binary feature records if the email is forwarded from another account to the recipient
3	Subj No Words	Records the total number of words in the subject line of the email
4	Subj Verify	This binary feature describes if the email's subject line contains the word verify
5	Subj Bank	This binary feature describes if the email's subject line contains the word bank

Table 2.12 Email body group

Id	Feature Name	Feature Description
1	Body html	Represents the presence of HTML in the email body
2	Attached file	Represents the presence if the email has an attached file
3	Body no function words	Total number of occurrences of function words in the email body such as account; access; bank; credit; click; identity
4	Body suspension	Binary feature represents the presence of the word suspension in the body of the email
5	Body verify/confirm your account	Binary feature represents the presence of the phrase verify your account in the body of the email
6	URL no IP addresses	Continuous feature that measures the number of links in an email that contain IP addresses rather than fully qualified domain names
7	URL at symbol	Binary feature represents the presence of links that contain an @ symbol
8	URL no links	Continuous numeric feature measures the number of links in the email body
9	URL no domains	Continuous feature that measures the total number of domains in all URLs in the email
10	URL max no periods	Continuous numeric feature measures the number of periods in the link with the highest number of periods
11	URL link text	Binary feature is true if the human-readable link text contains one or more of the following terms: click; here; login; or update.
12	Script scripts	Binary feature represents the presence of scripts in the email body.
13	Script java script	Represents the presence of java script in the email body.
14	Script popups	Binary feature that is true if the email contains pop-up window code.
15	Script no on click events	Continuous feature counts the number of on Click events in the email.
16	Invisible links	Represents if invisible links is found
17	Un matching URL	Describes if the visible URL is true
18	Long URL addresses	Describes if the email has many characters

2.5 Associative Classification (AC)

The main data mining categories are: supervised learning (predictive) and unsupervised learning (descriptive). In the former, the target is to construct a model that can perform one of data mining tasks such as classification. On the other hand, the target of unsupervised learning is to find patterns that condense the relationships

among items such as association rule mining. Moreover, predictive deals with one predefined target, while no predefined target in descriptive (Mishra et al., 2012).

Association rule mining finds frequent patterns, associations or causal structures among sets of items that have to pass a thresholds named, support and confidence, such as transaction databases where Market basket analysis is the famous example of association rules. On the other hand, classification is considered one of popular learning models in data mining. Its' target is to build a model to forecast unseen class through classifying database rows into a number of predefined classes. Common classifiers based on classification such as decision trees.Combine these two mining techniques, association rules and classification mining, brought out the associative classification (AC) approach, which first time was proposed by (Liu, et al., 1998) where the class attribute is the only attribute on the right hand side of the rule. Indeed, many studies such as (Kundu et al., 2009) had showed that AC often builds more accurate classification systems than traditional classification techniques. Moreover, differently from other classifiers such as neural network and probabilistic, which generate models that are not easy to understand by end-user, while AC generate "If-Then" rules that are easy to understand.

AC algorithms have two main phases: rule generation and classifier building, which predicts the class labels of all instances in test dataset to evaluate the classifier based on different evaluation measures. The first phase is handled by two steps: frequent ruleitems discovery and rules generation. The former discovers the frequent itemsets that is used by association rules mining. This step needs to pass the training dataset more than one time so it is caused an expensive computation effect (Lim et al., 2000) when finding all frequent itemsets. Frequent ruleitems are represented as a frequent k-ruleitems, where k represents the current number of the pass. The rule is considered as a frequent ruleitems if its support passes the minimum support constraint. Other frequent ruleitems are found through the correlation among previous frequent ruleitems. Finally, after all frequent ruleitems are generated then all possible rules that pass the minimum confidence constraint extracted from them and are used to predict the test data cases to evaluate the classifier based on the evaluation measures (Thabtah et al., 2010).

2.6 Associative Classification Related Definitions

As mention earlier a new approach that merges association rule with classification has first appeared in (Liu, 1998). Many experimental studies showed that AC is a high conceivable technique that develops more predictive and accurate classification systems than traditional classification methods like decision tree (Thabtah et al., 2005; Abu-nimeh et al., 2009). This is axiomatic while AC finds hidden correlations among the different features. Moreover, many of the rules found by AC methods cannot be found by other classification techniques (Soni et al., 2010). It is the process of deducing a set of class association rules Rs that pass predefined constraints (support and confidence) threshold to build a model to predict the class label of new instance. (Zhu et al., 2012; Liu et al., 1998) presented some related definitions to deal with AC.

Let D be the training data set with n attributes (*columns*) A_1, A_2, … ,A_n and D rows. Let C be a list of class labels. Specific values of attribute A_i and class C will be lower case a and c, respectively.

Definition 1: An item, or condition is defined as a set of attributes A_i together with a specific values a_i for each attribute in the set, denoted < (A_{i1}, a_{i1}), (A_{i2}, a_{i2}), … (A_{im}, a_{im})>.

Definition 2: The item support *itemsupp*(*i*) of an item *i* in *D* is the number of rows in *D* that contain *i*.

Definition 3: The support count *suppcount*(*r*) of *r* is the number of rows in *D* that matches *r's* condition, and belongs to *r's* class.

Definition 4: A rule *r* maps an item to a specific class label, denoted: <(A_{i1}, a_{i1}), (A_{i2}, a_{i2}),…, (A_{im}, a_{im})>*C*.

Definition 5: The rule support (*rulesupp*) of *r* is defined as the *SuppCount*(*r*)/*D*.

Definition 6: The actual occurrence (*ActOccr*) of a ruleitem r in T is the number of rows in T that match r's itemset.

Definition 7: The rule confidence (*ruleconf*) of r is defined as *SuppCount(r)/ActOcc(r)*.

Definition 8: Minimum support (*minsupp*) represents a threshold which discriminates among items that can be part of a rule (frequent) and

others that will be deleted. It is inputted by the user and normally fails between 2% and 5% according to (Liu, et al., 1998).

Definition 9: Minimum confidence (*minconf*) represents a threshold and it is inputted by the user which discriminates among rules where the strong rule is the rule that passing the *minconf*, so frequent itemsets and the *minconf* constraint are used to form rules.

Definition 10: An item in the training data set with *rulesupp* greater than the *minsupp* is known as a *frequent* item.

Learning (Training): The process to deduce Rs (knowledge) that pass the *minsupp* and *minconf* predefined values.

i. **Model (Classifier):** is when the prediction phase takes advantage of deduced rules Rs in the step above. It normally contains sequence steps, input data after pre-process step, learning, classifier handling rule ranking and pruning. Lastly, predicting step as a result for the model.

ii. **Itemset:** is a set of attributes together with their specific values for each attribute in the dataset, i.e<(X,x1)(Y,y1)> where X and Y are attributes and x and y are their values respectively.

iii. **Ruleitem:** is a set of itemsets with their class label. Such as < ((X, x1) (Y, y1)), (C, c1)> where c is the value of the class label C.

iv. **Classification accuracy**: is the number of cases where the predicted class of each test data matches actual class of test case for all cases in the test data.

v. **Training data:** is used to fit a model that can be used to predict a "response value" from one or more "predictors." The fitting can include both variable selection and parameter estimation.

vi. **Test data:** A data has the same training data characteristics, is used to test the accuracy of the model.

2.7 Common AC Algorithms

The majority of AC algorithms are based on two steps. Firstly, the generation phase uses association rule mining approach to find all possible rules that could be found in a dataset. This phase generates high number of rules which may cause redundant and over-fitting problems. On the other hand, the classifier builder is the second phase by AC algorithmsthat uses different techniques to prune redundant rule that are led to decrease the classifier accuracy rate. Many algorithms are built to deal with AC approach using different techniques in different phases and steps to solve problems coming with this approach such as the high number of generating rules. In this section we propose the common AC algorithms.

- **Classification Based on Associations (CBA):** The CBA is the first AC algorithm based on affinity analysis. It has two steps, a rule generator (called CBA-RG), which using the Apriori algorithm while discovering frequent ruleitems which needs more than one pass to training dataset. On the other hand, the classifier builder step (called CBA-CB), which is the beneficiary from the first step. The algorithm produces all the association rules depends on certain support and confidence constraints as candidate rules. CBA computes the support of a rule to know is it frequent or not in the first pass over the training dataset. Then in next subsequent pass it begins from rules already are found as frequent in the past pass to produce new possibly frequent rules called the candidate rules. These rules are considered as frequent and produced the rules (CARs) when the pass is finished. Also these CARs are built the classifier by the CBA-CB algorithm. CBA realizes two main conditions firstly; each training case is surly covered by the rule with the highest precedence between all rules that can handle the instance. Secondly, the rule is chosen when it is correctly classifies at least one remaining training instance. This algorithm is uncomplicated, but the many passes over the dataset let it inoperative (Gambhir et al., 2012). Moreover, it produces large number of rules which cause a computational load. Also it depends on single rule in prediction phase which may a weak rule.
- **Multi-class classification based on association Rule (MCAR):** (Thabtah et al., 2005) which is implemented based on Tid-list intersections to find the

rule as fast as possible. This algorithm contains of two main phases, rules generation and a classifier builder. Firstly, it needs just one pass over the training dataset discover the potential rules that has just one attribute, and then it intersects the potential rules Tid-lists of one attribute to find potential rules of two attributes and so forth. This technique is displayed as the first time in AC by MCAR which it requires just one pass over the training dataset. Secondly, the classifier is built based on rules that are produced in the previous step through test the efficiency of them over the training dataset where the potential rules that can handle certain number of training instances is used in the final classifier. Moreover, this algorithm adds a new rule sorting technique over previous ones such as confidence, support and rule length. MCAR considers the class distribution frequencies in the training data and choose rules that are connected with dominant classes. This new rule sorting techniques decrease the arbitrary selection of rules into the ranking step in different experimental results (Thabtah et al., 2008).

- **Classification based on Multiple Association Rules (CMAR):** thisalgorithm isbased on multiple association rules since some studies are proposed that AC suffers from the large set of mined rules and could be biased classification or has redundant rules while it is based on only one high confidence rule. CMAR is performed based on a weighted $\chi 2$ analysis using multiple strong association rules. It is highly effective and scalable (Li et al., 2001). Some AC algorithms is not easy to identify the most effective rule at classifying a new case such as CBA also a training data set usually produces a large number of rules. Dealing with these problems CMAR classifies the test instance using multiple rules instead of using one single rule in predicting phase. It deals with bias problem based on weighted $\chi 2$, which offers a good measure upon two thresholds support and class distribution for rule strength. It consists of two phases: rule generation and classification. It bases on a subset of high quality rules for predicting phase through analyzing them. If they give the same class label then it is classified the test case. Finally, CMAR acquires high accuracy than C4.5 and CBA (Gambhir et al., 2012). However, it is very slow because the using of FP Growth in generating rule phase and the overall accuracy rate can be improved.

- **Class Based Associative Classification Algorithm (CACA):** The majority of AC algorithms normally have three phases, Rule Generation, Building Classifier and prediction. First phase depends on the association rule mining approach to discover the frequent ruleitems. In the second phase rule sorting is handled the useful rules in a reasonable sort as well as deleting redundant rules. However, CACA algorithm merges both first and second phases together. As AC provides redundant rules hold in the classifier which increases the time cost when classifying test instance. Therefore, CACA comes with four new ideas. Firstly, it reduces the searching space of frequent pattern depending on class based strategic. Secondly, it offers a new structure named "Ordered Rule-Tree" to handle everything about the rules that prepare for the synchronization of the two steps. Thirdly, the compact classifier is unique and does not effect by rule reduction through redefine the compact set. Fourth, combine the rule generation and building classifier phase (Tang et al., 2007). Moreover, CACA is used in the association rules production the Apriori algorithm.
- **Associative Classification Based on Closed Frequent Itemsets (ACCF)**:A new classification method based on association was proposed to combat the challenging of the large set of rules that usually a training dataset find which are redundant during the classifier building (Li et al., 2008). The algorithm is based on the concept of closed pattern and extended an effective closed frequent pattern mining method, CHARM (Zaki et al., 2002). This method can be orders of magnitude smaller than the whole set of frequent itemsets while it derives all frequent itemsets. ACCF enhances the prune step for the redundant rules and the predicting method to classify a test instance. It trained on 18 databases form University of California Irvine (UCI) machine learning database repository which show that the algorithm produces a smaller number of rules, but without impact the classification accuracy rate when compares with the CBA algorithm.

2.8 Chapter Summary

In this Chapter we have reviewed the popular phishy emails methods which are used by machine learning techniques. Two methods were presented; traditional and automated each method has different techniques were used. Traditional methods

handle blacklist filters over password filters. On the other hands, automated methods handle statistical based methodssuch as Bayesian filter and Multi-layer methodssuch asneural network classifier. Compression among the previous techniques was proposed the advantages and disadvantages for each technique. Moreover, different phishing detections tools were presented at network level and server side filter with their advantages and disadvantages. Also we presented the popular research works handle this problem depend on the methods mentioned above and focusing on the features were used to analysis them in Chapter 3. Then common AC algorithms were presented and discussed focusing in some AC problem which we work on them in this thesis. Next Chapter presents our proposed model which we cover and deal with the disadvantages were found in different techniques and tools such as redundant rules, body content, fixed rules in the classifier by proposed new algorithm in different steps.

Chapter Three

The Proposed Model

3.1Introduction

Phishing email is an email fraud manner in which the scammer sends out legitimate-looking email (Almomani et al., 2012). In fact, all what the scammer needs just to collect personal, confidential and financial information from recipients. This information is used for identity theft(Ramanathan et al., 2012). As a result, the message appears to come from well-known and trustworthy banks or companies. However, legitimate businesses do not request any sensitive information through insecure channels (Alseadoon et al., 2012).Social engineering is what phishers use and depend on beside the number of different emails spoofing ruse to lure their victims.

In this thesis, we focus on AC data mining approach to handle the phishy email problem. As mentioned earlier, AC is a combined approach of two data mining methods, association rule and classification. The idea is to form a unity between them to construct classification models which offer easy yet accurate and understandable rules for the end users.

Classification is a learning function that categories data instances into one or more of several predefined categories. The data from which a classification functions or model is learned is known as the training set. In AC approach the training set is used by the algorithm to produce rules that are limited only to a particular attribute called the class in their consequent (European et al, 2007). Moreover, classification models could be used to identify many real world problems such as loan applicants as low, medium, or high credit risks so it properly predict the target class for each unseen instance.

In this chapter, we propose an enhanced AC mining algorithm based on Multi-class Classification based on Association Rules MCAR algorithm (Thabtah et al., 2005) and apply it on the hard problem of predicting phishy email. As a result, a new algorithm is outputted to Save Cyber from Phishy Email (SCPE) as one of the new web security models. SCPE is processed and trained to classify test instances in the phishing emails problem. This is performed on dataset consisting of significant features related to email that has been collected. This chapter structured as follows, the proposed model comes firstly in Section 3.2 then we focus on all phases that the algorithm performs and explain them in details. In Section 3.3 the feature

assessmentis handled. Rule learning, ranking and pruning are covered in Sections 3.4, 3.5 and 3.6 respectively. Section 3.7 explains the classifier building phase which is followed by prediction phase in Section 3.8. Finally the chapter summary is handled by Section 3.9.

3.2 The Proposed Model

The proposed detection model is shown in Figure 3.1. A collection of both, phishy and legitimate emails are used to obtain the significant features belonging to the phishy and legitimate emails permitting us to focus on them in the training step. The feature assessment phase is handled in Section 3.3, because it offers us the input for the SCPE algorithm which is forming dataset having emails. On the other hand, two thresholds must be inputted named minimum support and minimum confidence. The proposed algorithm is trained on the dataset and using the specified thresholds to generate the knowledge (rules). Differently from other AC algorithms SCPE scan the dataset one time only by employing an intersection method based on Tid-list to enumerate the location, support and confidence of mining items inside the training dataset. The Tid-list offers a representation of the dataset having all necessary information related to each item (attribute value). Figure 3.2 presents the learning step in Figure 3.1 and shows how it finds potential rules depending on the above support and confidence values. Firstly the algorithm discretizes continuous attributes (Section 3.3 covers this step). Next, it produces frequent one-ruleitems (F1) which has just one attribute. After that it combines ruleitems conditions by intersecting their set of Tid-list to produce candidate ruleitems involving more attributes (F2), (F3) and so forth. However, if any new rule does not pass the thresholds it is deleted immediately (Section 3.4 illustrates this step in details). The proposed algorithm scans the database exactly once reducing I/O costs. As a result, intersecting the Tid-list using vertical representation does not suffer from any of the overheads problem (Zaki et al., 1997).

The knowledge discovered by an AC algorithm is called “Class Association Rules” (CARs). Lastly, in Figure 3.1 the proposed model builds a classifier based on the produced CARs. We introduce new techniques to improve two phases in the life cycle of any AC algorithm. These are the classifier building and prediction. In the classifier building (Section 3.7), a new ranking technique is given the superior rule the highest rank over all other rules; such rule is a rule has the maximum number of attributes in its antecedent side then we use the general ranking techniques Section 3.5

handles this step in details. As well as a new rules pruning technique are exhibited. A partial matching is proposed in rule valuation to let the number of remaining rules in the classifier smaller than current AC algorithms such as CBA or MCAR. At this point we solved one main problem associated with AC algorithm by cutting down the classifier size after implementing the new rule pruning procedure, Section 3.6 handles this step. Finally, the prediction phase which is handled in Section 3.8 takes advantage of new technique that takes into account more than one rule instead one single rule in predicting (Phishy or legitimate) test data.

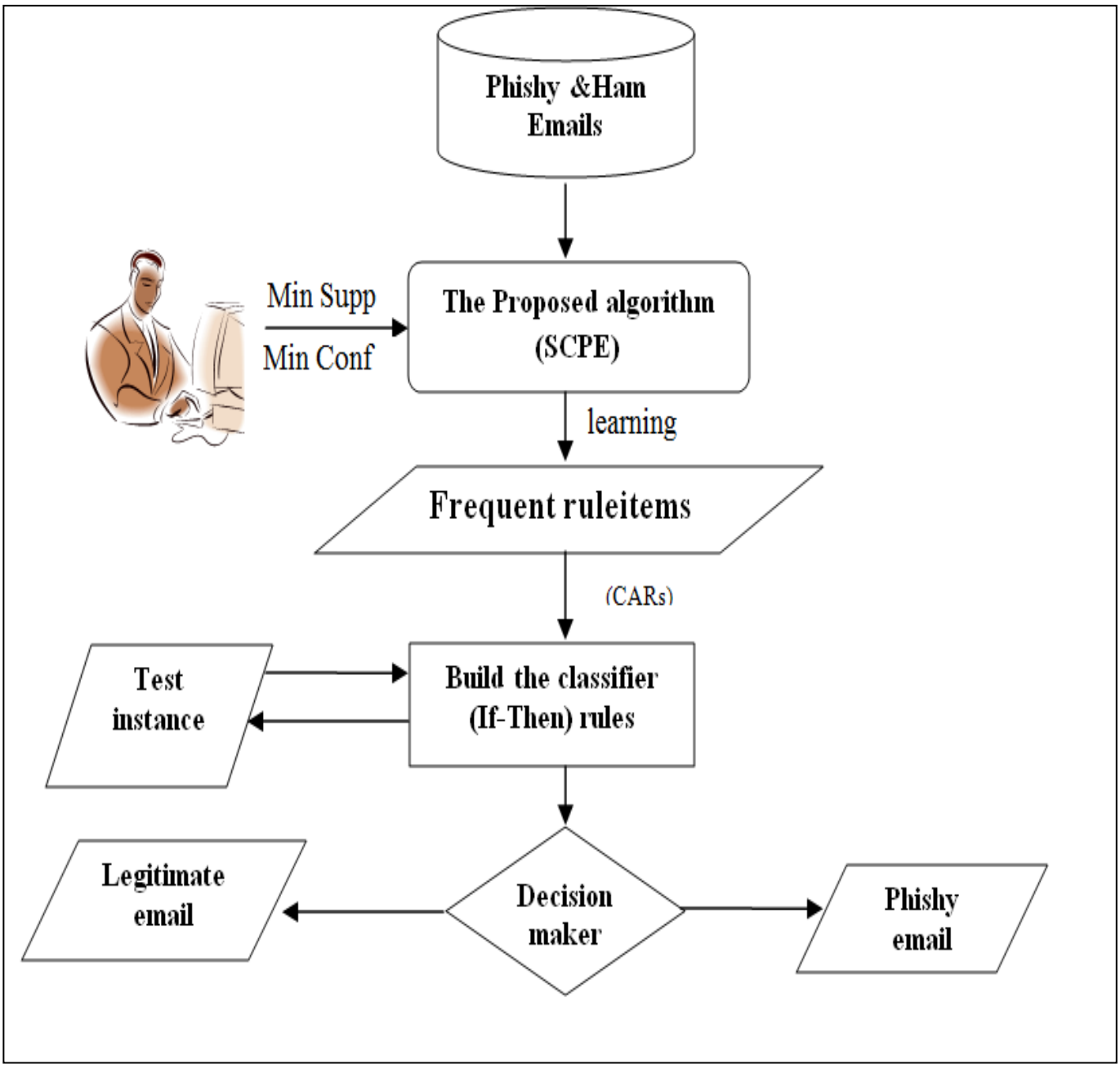

Figure 3.1: The proposed model of phishing email

Input: Training data (T), minsupp and minconf thresholds
Output: A set of CARs

Preprocessing phase
Discretise continuous columns

The Algorithm
Scan T for the set R of frequent one attribute-value
Do
For each pair of disjoint items V_1, V_2 in R
Intersect the sets of rowIds of $V1$, and V_2 and store it in Ts
If Ts size <itemsupp then
prune the new item
else
begin
If ($<V_1 \cup V_2>$, c_i) passes the minsupp threshold
begin
if ($<V_1 \cup V_2>$, c_i) passes the minconf threshold
begin
Generate a rule for $<V_1 \cup V_2>$ if it passes
$R \leftarrow R \cup <V_1 \cup V_2>$
end if
else discard the new item.
end if
endif
end

Figure 3.2: SCPE learning algorithm

3.3 Feature Assessment

In this assessment, the email features are used to classify the type of emails (phishy or legitimate). These features are divided into two types “nominal” and “continuous”. Therefore, according to Figure 3.2,which shows the discrete step is done only for the continuous features to be nominal one since we are dealing with classification dataset. This has been performed using the multi-interval discretisation technique (Fayyad et al., 1993) in WEKA software. Hereunder, we briefly explain the discretisation of numeric attributes. Firstly, the continuous attribute is sorted in ascending order with the class values associated with the instance belonging to it. Then, breaking points is placed whenever the class value changes to calculate the information gain for each possible breaking point. The information gain represents the amount of information claimed to an attribute value with respect to its gain. Finally, the breaking point that minimizes the information gain over all possible breaking

points is selected and the algorithm is triggered again on the lower range of that attribute.

Email features are the main criterion to determine the class type of a test instance. As a consequence of that, 23 featureswere categorized into two groups: email header group containing five features and email body group involving eighteen features presented in Tables 2.11 and 2.12 in Chapter two.theanalysisof feature depends on their frequency (Habib et al., 2012) to specify the number of times a feature appeared in the dataset, indicating which features areof high repetition. The Count function from Oracle 11g is used in SQL query to get features frequencies. This function is an aggregate function that counts the number of rows accessed in an expression allowing on all types of expressions (Nikolov, 2011).The dataset was loaded on Oracle database using Oracle Database Container (ODBC). Each attribute is grouped with the class in respect to four cases, if it appears or not with the two type of classes (phishy or legitimate). The result of the SQL statement is displayedin Table 3.1 which shows the appearance of the feature depending on the class with its frequency in the whole dataset where '1' means appear and '0' otherwise. On the other hand, continuous attributes are handled by discretisation technique as described earlier. Features analysisbrought out new feature groups for the proposed model. Feature's data type was the main criteria to group them. As a result, two new groups are identified, Binary features group and Continuous features group.

Amazingly, none of the header group features that are shown in Figure 3.3 were chosen. This elimination is fastened on their frequency analysis in the dataset even if they are used frequently in many classifiers. The first two features, "SubjReply" and "SubjForward" as observed are not frequent at all so we ignored them and did not give them any weight. Moreover, we found the majority of emails send by their owners directly, so neither "reply" nor "forward" features are used. On the other hand, the next two features "SubjVerify" and "SubjBank" were encountered as not frequent as well. However, we have not ignored them because they have a logical weight since banks never ask their agents for any confidential information such as passwords via emails (Alseadoon et al., 2012). Obviously we left them in a new feature which will be discussed below because any email has them should be taken as a seriously disposal.

Table 3.1: Sample of feature frequency analysis function result for “SubjReply”

APPEARANCE	Class	Frequency
‘1’	Phishy	44
‘0’	Phishy	656
‘1’	Legitimate (Ham)	89
‘0’	Legitimate (Ham)	211

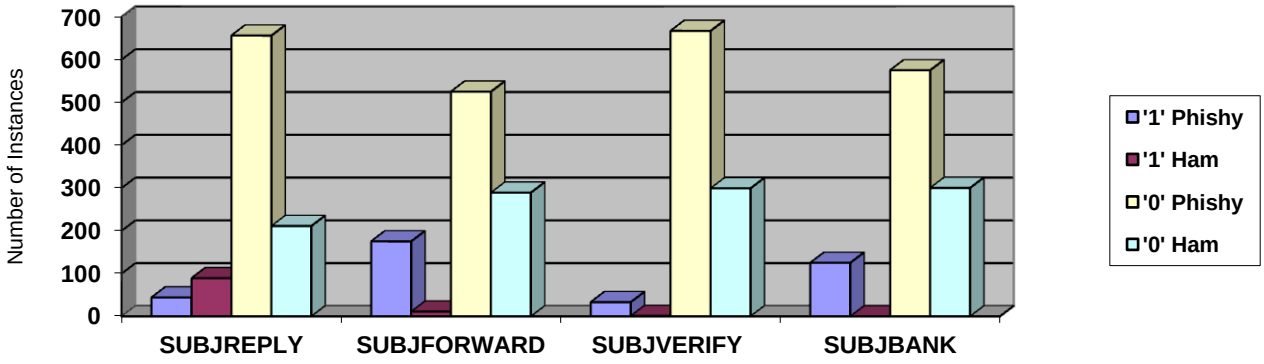

Figure 3.3: Email header feature frequency analysis

The last feature in email header group is “SubjNoWords” which is the only continues feature in the group. Its weight was based on the average value in Phishy and legitimate emails. In other words, we have checked how many words in average in the subject line in both types using the aggregate average function respect to the class attribute by Oracle SQL statement. We concluded with 5.04,5.26 for phishy and legitimate respectively. Since the percentage of phishy emails in our dataset is 70% to 30% for legitimate emails so we ignored this feature and did not consider it, because has not discriminated among classes. Now let us analyses and discuss the second group of features: The email body group. Starting with two frequent features as shown in Figure 3.4, they are “BodyHtml” and “BodyFunctionsWords”. These two often show up with phishy emails so they were hired. However, “BodyFunctionsWords” feature has been added to binary feature group and named it “EmailFunctionWords” even though it was identified as continuous feature. This let us cut down the number of features by including them in it from both header and body of the email. Table 3.2 shows the function words we choose to use in this feature. The following features, “BodySuspicious” and “BodyVerify/Confirm” are not frequent but we found a

relation between them and the previous two header group, ("SubjVerify" and "SubjBank") so we included them in our feature set "EmailFunctionWords".

Table 3.2: list words of EmailFunctionWord feature

Award	Business	Account	Bank	Click
Credit	Access	Update	Identity	Validate
Card	ATM	Congratulation	Amount	Money
Prize	Gain	Win/Won	Financial	Fund
Loan	Lottery	Claim	Payment	$/Bound

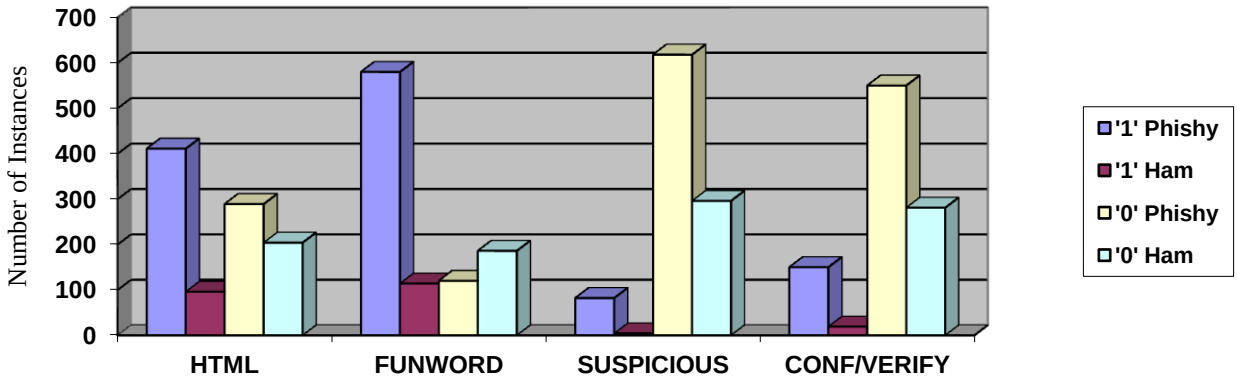

Figure 3.4: Part of body feature frequency analysis

Following the email body group we have found two other features, "UrlNoIpAddresses" and "UrlAtSymbol" that are not frequent, so we ignored them. Moreover, phishy email features and phishy websites features overlap in certain features such as "LongUrl". However, some features operating seamlessly with phishy websites rather than phishy emails from the body group, "ScriptScripts", "ScriptJavaScript" and "ScriptPopups". One of these features is usually come with phishy website, while in phishy email are not frequent at all so we have avoided them from our selection.

Further, "UrlNoLinks", "UrlNoDomains", "Invisiblelinks" and "UrlMaxNoPeriods" are continuous features in the second group except the "Invisiblelinks" which is binary one. This set is significant and effective in phishy email since they are frequently found so we consider them. Phishy emails often have large number of links that lure the user while they are dependent on different fraud domains and invisible links which appear as a legitimate link but in fact a hidden links are handled by them. However, "UrlMaxNoPeriods" is omitted since we handle it in

other feature called “LongUrl”. Figure 3.5 shows the last four features. The first two, “UnmatchingUrl” and “LongUrlAddresses” are much related with phishy emails their frequencies in the dataset are high. They take an important role to lure users so they are chosen. On the other hand, the last two, “AttachedFile” and “OnClick” are not that frequent but when we take each one in particular, “Attached File” it has never been used before we have suggested to select it and add it into our group, since it could be an effective feature if it is attached with malicious software (Activity, 2012). Finally, “OnClick” feature has not that much occurred with legitimate email so we have added it under the new name “EmailFunctionWords feature”.

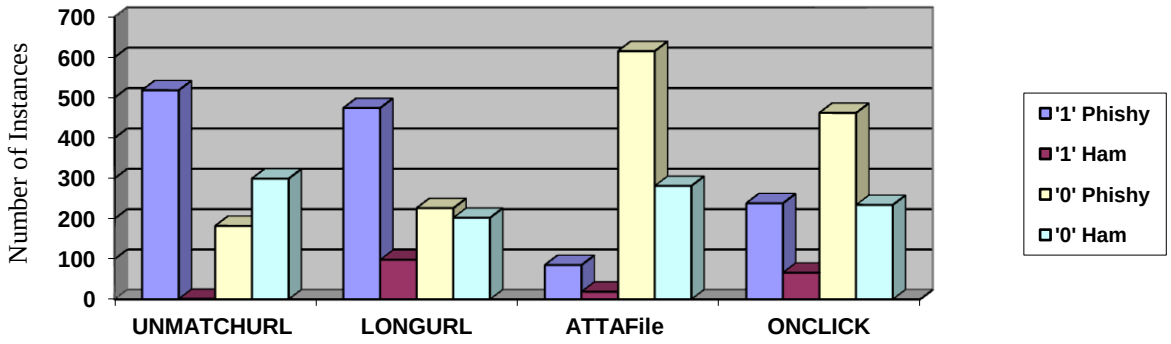

Figure 3.5: Frequency analysis last fouremail body feature

Table, 3.3 & 3.4 show the binary and continuous features respectively that we have used in our experiments. These features deem the most significant ones to deal with in phishy email problem. On the other hand, Table 3.5 shows the features that have been eliminated.

Table 3.3: Binary features

Feature Id	Feature Name	Feature Description
1	Body Html	Represents the presence of HTML in the email body
2	EmailFunctionWords	Describes if function words in the email header and body such as account; suspicious; bank; verify; click; see Table 3.2
3	InvisibleLinks	Total number of invisible links
4	UrlLinkText	If the human-readable link text contains one or more of the following terms: here; login; or update.
5	UnmatchingUrl	Describes if the visible URL is true.
6	LonUrlAddresses	Describes if the email has many characters.

Table 3.4: continuous features

Feature Id	Feature Name	Feature Description
1	UrlNoLinks	Feature measures the number of links in the email body
2	UrlNoDomains	Measures the total number of domains in all URLs in the email

Table 3.5: Eliminated features

Feature Id	Feature Name	Ignored Reason
1	SubjReply	Not Frequent
2	SubjForward	Not Frequent
3	SubjVerify	Included in feature number 2 in Table 3.3
4	SubjBank	Included in feature number 2 in Table 3.3
5	SubjNoWords	Does not discriminate the different classes
6	BodySuspension	Included in feature number 2 in Table 3.3
7	BodyVerify/Confirm	Included in feature number 2 in Table 3.3
8	UrlNoIpAddresses	Not Frequent
9	UrlAtSymbol	Not Frequent
10	UrlMaxNoPeriods	Included in feature number 6 in Table 3.3
11	ScriptScripts	Phishing Websites Feature
12	ScriptJavaScript	Phishing Websites Feature
13	ScriptPopups	Phishing Websites Feature
14	OnClick	Included in feature number 2 in Table 3.3
15	Attached file	Included in feature number 2 in Table 3.3

3.4 Rule Learning

Our algorithm is a special case of association rule that considers only the class label as a consequent of a rule to deduce a set of class association rules (CARs) from the training dataset which satisfy certain user-constraints, minimum (support and confidence) thresholds. The majority of AC algorithms such as, CBA2 and CMAR require multiple scans on the dataset in order to discover the knowledge(Mahmood et al., 2007).However, SCPE goes over the training dataset only once to count the occurrence (support) of one-ruleitems and Tid-lists (sequence) are stored in a vertical

format, while the ruleitems that do not pass the minimumsupport are eliminated. Next, intersecting the sequence of two disjoint one-ruleitems is used to generate the candidate two-ruleitem and so on until all frequent ruleitems are found.

This learning method has been used before in association rule by (Zaki et al., 1997). It transforms the training dataset into items table containing the sequence of each item in the training dataset, and then it employs simple intersections among these sequences to discover frequent values and produce the rules. Since this approach iterates over the training dataset only one time therefore it is efficient with regards to processing time and memory utilization. Table 3.6 shows part of training data and the next explanation illustrate the rule generation phase using the vertical technique with sequences. To show how we determine a frequent ruleitem, consider for instance itemsets in Table 3.6 < (Body Html, *1*)> and < (EmailFunctionWords, *1*)>. The next two sets represent the sequences in which they occur, {1, 3, 4, 5, 6, 9, 10} and {1, 2, 4, 5, 8, 9, 10}. We can determine the support of the itemset< (Body Html, *1*)>, < (EmailFunctionWords, *1*)> by intersecting the sequences sets for itemsets< (Body Html, *1*)> and < (EmailFunctionWords, *1*)>. The cardinality of the resulting set {1, 4, 5, 9, 10} represents thesupportfor itemset< (Body Html, *1*)>, < (EmailFunctionWords, *1*)>, i.e. 5/10. If it passes the minimum supportthreshold, then we proceed by checking whether there is some class C such that < (Body Html, *1*)>, < (EmailFunctionWords, *1*)>C> passes the minimum supportthreshold, otherwise we prune it.

Table 3.6: Part of training data

Sequence	Body Html	EmailFunctionWords	Class
1	1	1	Phishy
2	0	1	Legitimate
3	1	0	Phishy
4	1	1	Phishy
5	1	1	Phishy
6	1	0	Legitimate
7	0	0	Phishy
8	0	1	Legitimate
9	1	1	Legitimate
10	1	1	Phishy

The support and confidence for a frequent ruleitem is calculated by our algorithm by locating the largest class that appears with ruleitem as we will discuss below. Then by taking the cardinality of the set where the ruleitem and its largest class occur and dividing it by the size of the training dataset, the support of ruleitem is obtained.

The confidence similarly is calculated as support except that the denominator of the fraction is the size of the set of the last sequence of the ruleitemcondition (itsitemset) instead of the size of the whole training dataset. Frequent ruleitems are generated recursively from ruleitems conditions having a smaller number of attributes, starting from frequent one-ruleitems comes from a single pass through the training dataset. It should be noted that every time a frequent ruleitem is found, only the rule with the largest confidence is considered.

Consider the vertical data layout shown in Figure 3.6 for the training dataset of Table 3.6 as an example to illustrate the rule generation process. Assume that minimum supportand minimum confidencehave been set to 40% and 60%, respectively. These thresholds have been set only for example purpose. During the scan, the frequent one-itemsets that pass the minimum supportthreshold are identified, (Body Html, *1*), (EmailFunctionWords, *1*) and all other infrequent itemsets and their sequences are discarded. Candidate two-itemsets, which are produced by merging disjoint frequent one-itemsets are shown in bold in table 3.7. Once these itemsets are identified, we check their supports and confidencessimultaneously bylocating classes that occur with their sequences.

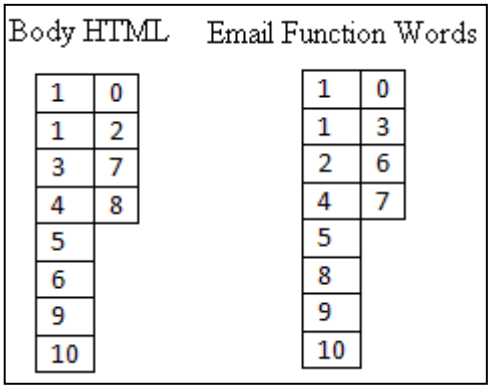

Figure 3.6: Vertical data representation for the training data

For example, for candidate two itemset< (Body Html, *1*) (EmailFunctionWords, *1*) > we locate its classes using its sequence {1, 4, 5, 9, 10}.

We choose the class with the largest frequency, which is Phishy, and divide the cardinality of the sets {1, 4, 5, 10} by the size of the training dataset, to obtain the support for ruleitem<((Body Html, *1*) (EmailFunctionWords, *1*), Phishy>. If it has enough support, we calculate its confidence by dividing the size of sequences of the ruleitem's largest class, i.e. 4, withthe size of the ruleitem'scondition set, i.e. 5. For ruleitem< (Body Html, *1*) (EmailFunctionWords, *1*), Phishy>, the support is 4/10 and the confidence is 4/5. In the case that the ruleitem passes theminimum confidence threshold, we immediately add it as a potential rule in the classifier. Otherwise the rule is pruned.

Moreover, there is no separate phase to calculate the confidences for all frequent ruleitems in SCPE, whereas the majority of current AC techniques produce frequent ruleitems in one step and find their confidences in a separate step.

Table 3.7: Possible frequent two-itemsets generated from Table 3.6

(1, 1)	**(1, 0)**	**(0, 1)**	**(0, 0)**
1	3	2	7
4	6	8	
5			
9			
10			

3.5 Rule Ranking

Ranking of generated rules reflects the strength of the classifier since in this step we use the selected rules to predict test instance in later phase. For example, the majority of AC algorithms rank the rules in respect to the confidence and support levels. When several rules have identical confidences and supports, they arbitrary choose one of the rules which could decreases the accuracy rate of the classifier. Therefore, SCPE usually focuses on the superior rule in the final classifier. The superior rules are ones with large number of attribute not only the ones with high confidence values against the training dataset.

Our contribution in ranking is to deal with preferring superior rules that have maximum number of attributes. So our algorithm prefers specific long rules over general short rules in antecedent side. This is since specific rules are more accurate in

predicting test instance especially because they cover smaller number of training cases. If two or more rules have similar number of attributes values in their antecedent, then we choose the rule that has largest confidence. If the confidences are the same we prefer the rule with the largest support.

So we follow the following procedure when two or more rules having same length: R1 > R2 if the confidence of R1 is greater than R2. When the confidence values of R1 and R2 are the same, but the support of R1 is greater than R2 the algorithm tend to favor R1. R1 precede rule R2 In the case of confidence and support values for both of them are the same, R1 is generated before R2 so R1 is preferred. As a result, there is minimal chance to choose any rules randomly which may led increasing classification error rate.

3.6 Rule Pruning

Association rule mining considers the correlations among all attributes in the training dataset (Chen et al., 2005) and therefore, the produced rules may overlap in their training instances. After rules ranking the pruning will start to choose only effective rules in the classifier. Now, the rules are ranked, then starting with superior rules if it covers at least one training instance it will be inputted into the classifier. The rule is pruned when it fails to classify at least a single instance. This way the algorithm is eliminating any rules that are redundant or contribute to incorrect classification. Our algorithm applies partial matching as new criteria if full matching of the candidate rule body and the training set is not met. This technique let the classifier contains less number of rules because a rule now has more training instance coverage.

For example, Table 3.8 represents a training data set. Suppose we have the following rules: R1 – R3 as follow:

R1: ((Body_Html, 1),(EmailFunctionWords, 1),(UrlLinkText , 0), (UnmatchingUrl, 1) →Phishy).

R2: ((Body_Html, 1),(EmailFunctionWords, 0),(UrlLinkText , 0), (UnmatchingUrl, 0) →Legitimate).

R3: ((Body_Html, 0),(EmailFunctionWords, 0),(UrlLinkText , 0), (UnmatchingUrl, 1) →Phishy).

Table 3.8: Training DataSet

TID	Body Html	EmailFunctionWords	UrlLinkText	UnmatchingUrl	Class
1	1	1	0	1	Legitimate
2	1	0	0	1	Legitimate
3	0	1	1	0	Legitimate
4	1	0	0	1	Legitimate
5	0	1	1	0	Legitimate
6	1	0	0	1	Phishy
7	0	1	1	0	Phishy
8	1	1	0	1	Legitimate
9	1	1	0	1	Phishy
10	1	1	0	1	Phishy

Starting from R1, the body of R1 matches the values in TID (1, 8, 9 and10). Thus, R1 will be added to the model and these training instances are removed from the training dataset. After that, R2's body does not match any TID in the training dataset. Therefore, to decide whether R2 part of the model or not we look at the training data Table 3.8 we find three cases having three values corresponding to three items of R2 TID (2, 4 and 6) so R2 will be added to the classifier and all training data are removed. Then R3 has also three training cases in one item and as a result, R3 will be part of the classifier and all training cases will be deleted.

3.7 Classifier Builder

SCPE deal with a rule can cover at least one training instance. After rules are generated, Figure 3.7 presents the classifier builder algorithm used by SCPE:

Starting with the first rule ri, we fully match it with training rule if it classified a single rule at least, we simple delete its occurrences from training data (sequences) container, Tiand delete it from R' then put it into classifier Cl. Then For each other potential rule ri, we check if it partially covers at least one training instance into Tithen this rule will be inserted into the classifier. Finally, we choose a default class by subscribing the labels in Ti from C container which has all labels for the whole dataset or by take the majority class as a default class from the current Cl and add it to Cl as a default class if Ti is empty. Equation 3.1 represents the time complexity forthe classifier building phased which is used by SCPE.

$$\sum_{i=1}^{n}\sum_{j=1}^{k}(C1 + C2) \text{ where } n.k \leq n^2 \quad \sim \quad O(n^2) \tag{3.1}$$

According to equation 3.1 the first loop refers to the training instances in the dataset takes linear time, which the first summation symbol represents it. Consequently, the

second loop refers to the generated rules, which is nested inside the first loop takes also linear time where the second summation symbol represents it. These two loops add a dimension of n and k time complexity, where n and k are the numbers of runs for loops. Moreover, C1 and C2 are used as constants to represent the IF statements where C1 belongs to the first one which classifies the instance in fully matching and partial matching approaches if the full matching fails. On the other hand, the C2 refers to the second IF statement which deals with finding the default class.

```
Inputs: set of created rules (R), training container (Ti), class container (C)
A rule r in R has the following properties: Items, class, rowIds (tid-list)
The class container, C contains the occurrences of class labels in the training data
Output: classifier (Cl)

        For each training data do
        For each rule r ∈ R' in sequence do
        begin
            If ri classifies at least a single instance
        insertri in Cl
        delete all its coverd instances from Ti
        else
        ifri classifies partly at least a single instance
        insert r into Cl;
        delete the classified training from Ti
end if
end if;
ifTi.size> 0 then
select the majority class as a default class from (C-Ti)
else
select the majority class as a default class from the current Cl and add it to Cl
end if
end
```

Figure 3.7: SCPE classifier builder algorithm

3.8 Prediction of Test Data

In data mining, prediction is the process that forecast the class label for unseen instance which is ultimate goal of classification. To illustrate the idea, let R be the set of generated rules and Ts be the set of test data instance. Figure 3.8 shows the proposed test prediction method used in our algorithm. To classify a test instance, the proposed algorithm uses a simple method, which states that the first rule in the set of already ranked rules that contained in the test instance classifies it. If there is no rule fully matching the test instance, SCPE uses a new process to find all rules that match

part of the test instance and calculates the average confidence of rules that have the same class then applies the class of the high average of confidence. So the prediction here is based on mathematical formula (3.2) below where *conf* is the rule confidence and *ri* is the number of rules belonging to the same class.

$$\sum ri\ (conf)/ri \qquad (3.2)$$

This way we ensure the given class is dependent on a number of rules applicable to the test instance. Also more than one rule is playing a role to classify test instances which is surly better of using one rule as MCAR and CBA. In cases where no rule matches the test instance, the default class is assigned to the test instance.

```
Input: Classifier (R), container (Tr) and test data instances (Ts)
Output: Predicted class

Give a test case tc, the classification process works as follow:
  For each test instance Ts Do
  For each rule r in set of ranked rules R Do
    If r classifies tc
assigntc the class
else
  Find all applicable rules that partially match tc and store them in Tr
  If Tr size > 0
calculate average confidence of all rules belong to the same class
assign the class of the highest confidence average
else
assign the default class to tc
end if
end if
emptyTr
end
end
```

Figure 3.8: SCPE prediction algorithm

3.9 Chapter Summary

The proposed model is an AC approach that operates two steps (rule ranking, rule pruning). We have used a new ranking that prefers superior rules and then enhanced rule pruning by using partial match between the rule's body and the training instance. Moreover, the prediction phase is enhanced by using group of rules average confidence for predicting the test data instances. SCPE offers also vertical data format which let the scans over the training dataset to be only one time during learning rules differently from most AC algorithms. The classifier will be evaluated in next chapter

against phishing dataset according to different evaluation measures. Then a comparison is done with different classification algorithms.

Chapter Four

Data and Experimental Results

4.1Introduction

Different classification algorithms are compared with our algorithm according to classification accuracy, and number of rules for phishy and legitimate emails dataset. The main evaluation measures that are used in our comparison are precision and recall. Three resources were used to collect this dataset which is mentioned in Section 4.2. In order to evaluate our algorithm, it has been compared with three classification algorithms: The reason behind selecting these algorithms is the different training strategy they use in discovering the rules.

1. NaiveBayes filter belongs to statistical text classification systems. It has used in many fields of science; this theory depends on the previous event to prove the conditions and give the optimal solution to solve any problem (Domingos et al., 1997).
2. J48 classifier is among one of the most popular and powerful decision tree classifiers. It is the improved versions of C4.5 algorithms proposed by (Quinlan, 1993).
3. PRISM is one of the rule induction algorithms which can only deal with character attributes and does not do any pruning using coveragesearch. It implements a top down. Moreover, it is produced an illustration for the classification result as well as used directly for decision making (Romero et al., 2010).

The experiment was running on Intel Pentium 4 cor i3, 2.66 GHz, 2 GB RAM, Hard Disk 160 GB and Windows 7 workstationsbased on cross validation as a testing approach to build the classifiers of the proposed algorithm and thethree compared algorithms. This technique is separated the training dataset into (n+1) folds arbitrary. Then rules are learned from n folds and evaluated on the remaining hold out fold. The process is repeated n+1 times and the results are averaged and produced. The experiment was executed using 10 fold-cross validation (Yin et al., 2003). Moreover, Minimum Support and Minimum Confidence are 5% and 60% respectively in order to discover the highest number of rules, similar to other research studies, i.e(Liu, et al., 1998). The proposed algorithm has been implemented using Java, and the results of NaiveBayes, J48 and Prism were derived from WEKA which is also implemented

using Java. Figure 4.1, shows the interface screenfor it which is an open source machine learning tool (WEKA, 2002).

Figure 4.1 WEKA Interface (WEKA, 2002)

The classifiers of all algorithms were trained based on 1000 emails, 700 phishy and 300 legitimate were handled by Excel file. This file was converted to CSV format by save as the Excel file to this extension CSV. "CSV 2 ARFF" (Marko, 2012) is an online tool to convert CSV(spreadsheet) format files to ARFF WEKA format file. Also this tool is offered the user to choose the data type of the feature (attribute) to be known by the classifier, numeric or nominal during the converting period. Moreover, the eight features mentioned in Tables 3.3 and 3.4 were used as the attributes for the four algorithms see Figure 4.2 which presents them depending on WEKA software.

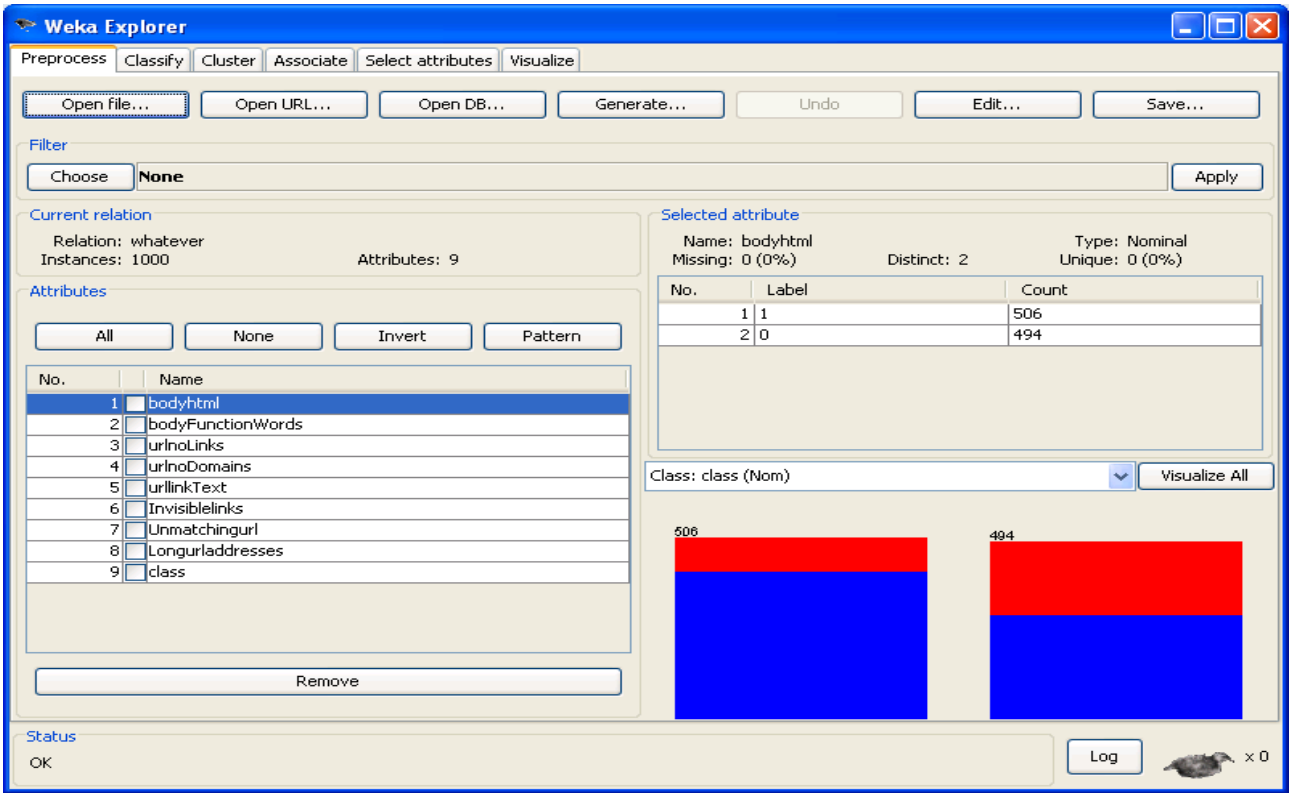

Figure 4.2 The Attributes used by experiments (WEKA, 2002)

4.2Dataset

The dataset of this work was a big challenge. Unlike the Spam emails datasets which are frequently found in many resources with a good explanation illustrate them. However, phishing dataset is hard to gather and analyses since there are not that enough resources to deal with. On one hand, the first part of the dataset is the phishy emails which we have depended on two resources to collect them and extract the whole features. We had to deal with it by extract the features manually without using any software because of the way that emails were stored which took long time and hard work. The first resource is Scamdex which is the first and foremost having a huge archive of scam emails since 2003. It covers many type of scam over the internet such as, 'Get Rich Quick', Bogus Lottery Winnings, Spoof Phishing Emails from Banks and all other Identity Theft and Internet Fraud (Scamdex, 2012). Millersmiles source has information about spoof email and phishing scams which is offered daily reports of new scams that is in circulation which is our second resource (Millersmiles, 2012). Also was originally founded on 2003 by Mat Bright. The story was started when he wanted to use the site to sell and promote book collecting. Buying and selling online let him comes across many dangers where the biggest of these was the threat from spoof email and phishing scams.

On the other hand, the legitimate emails were easy to collect than phishy emails. We depend on resources of Spam dataset the first is CSDMC2010 SPAM corpus dataset and Spam Assassin dataset **(**Spam-assassin, 2012**)** which are offers the legitimate datasets for the data mining competition. The later resource handle,legitimateemails that contains two categories: easy legitimate emails, and hard legitimate emails which is very close to spam.As a result, the model was trained by using them spatially we use AC technique which can find all relations between attributes.

4.3 Accuracy

Accuracy is the rate of correct predictions that the model achieving when compared with the actual classifications in the dataset. Figure 4.3 shows the accuracy of our classification model in predicting the phishy emails problem compared with three algorithms in Section 4.1. Figures 4.4, 4.5 and 4.6 show the accuracy result for each algorithm:NaiveBayes, J48 and Prismrespectively based on WEKA software. Accuracy distinct between the different filters techniques and find the best one of them to be used to avoid the phishy emails. In the phishy email problem, the accuracy is the phishing percentages that classify as phishing email and the legitimate percentages that classify as legitimate email. But the main problem occurs when the phishy email is classified as legitimate email, and the legitimate email is classified as phishing email. The main goal of phishing filtering is to solve these problems; Table 4.1 shows contingency table which can be constructed to resolve this problem.However, the proposed algorithm got accuracy rate close to J48, but the classifier based on our proposed model is easy to understand model and has a promising performance while it uses AC approach. Moreover, the average precision and recall result of the proposed model is better than result got by J48 which reflect the strength of the proposed model over J48 next section illustrate that.

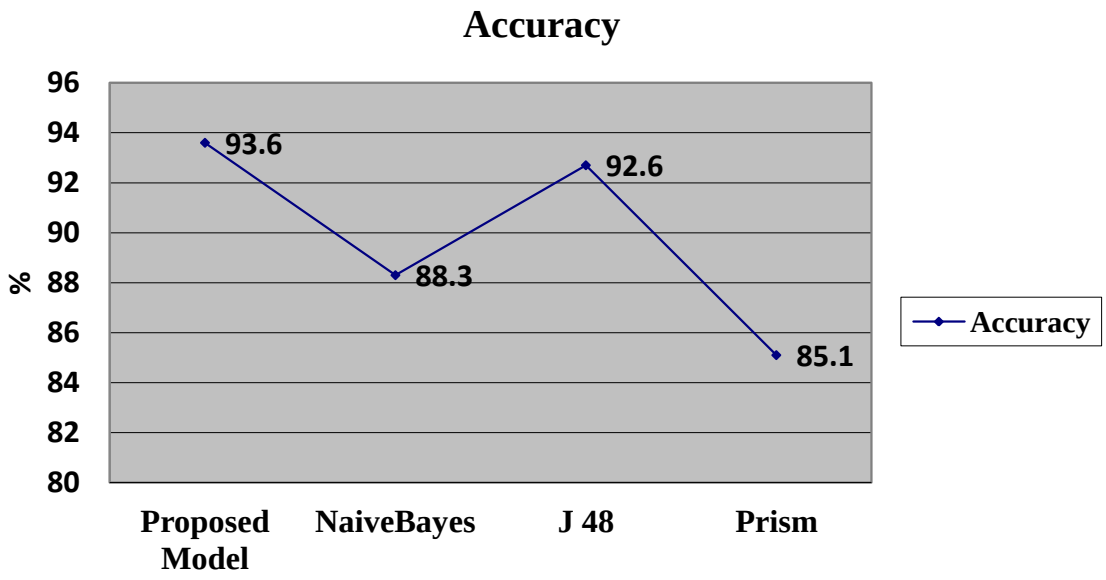

Figure 4.3: The Accuracy Result

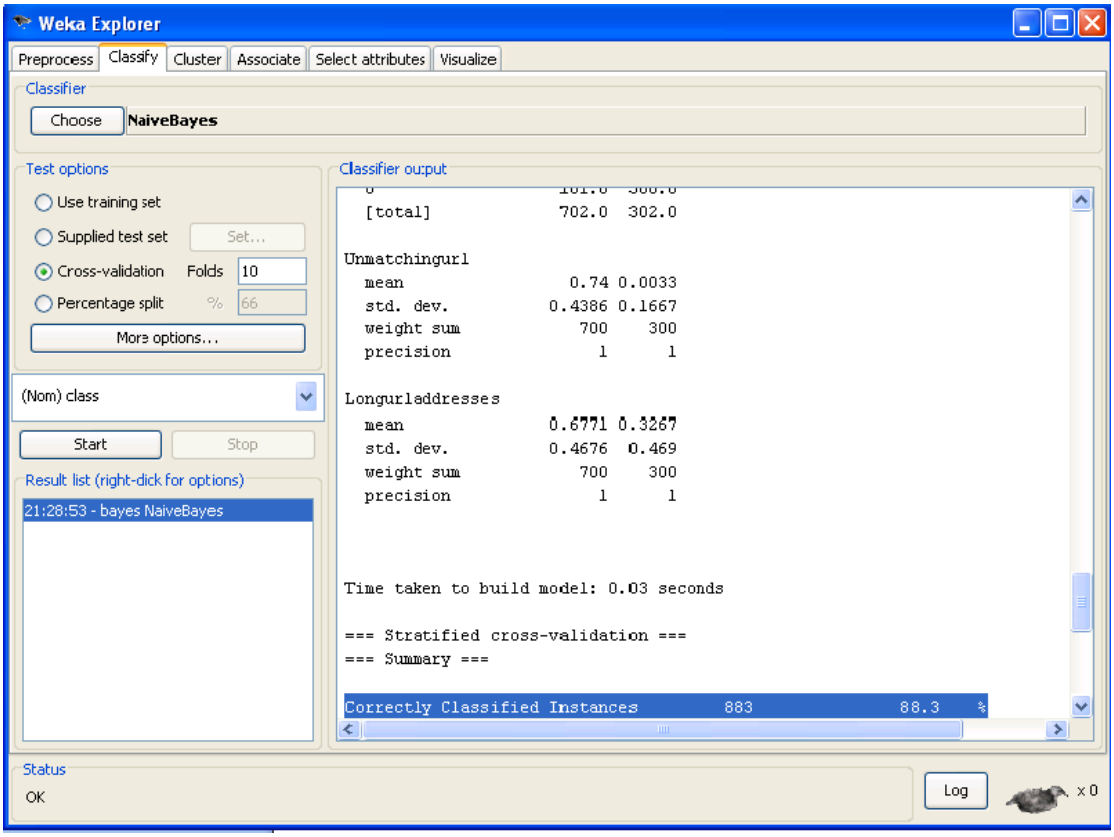

Figure 4.4The accuracy result of NaiveBayes algorithm (WEKA, 2002)

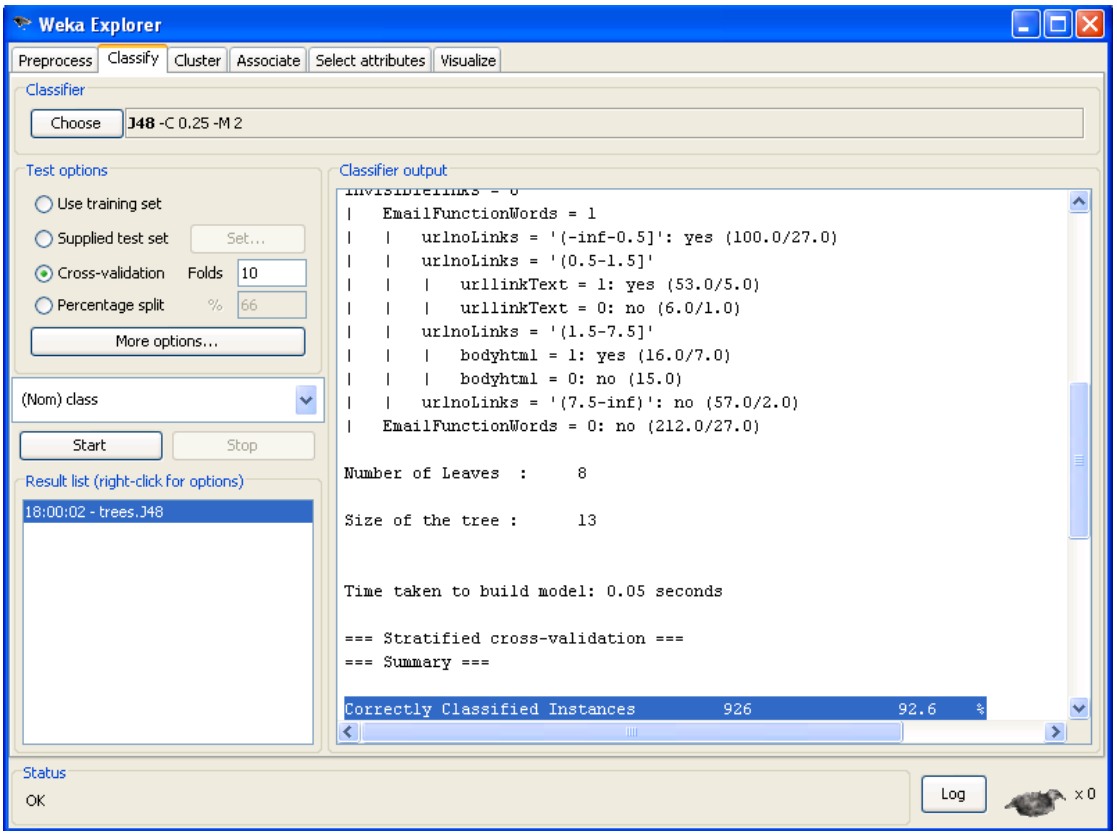

Figure 4.5The accuracy result of J48 algorithm (WEKA, 2002)

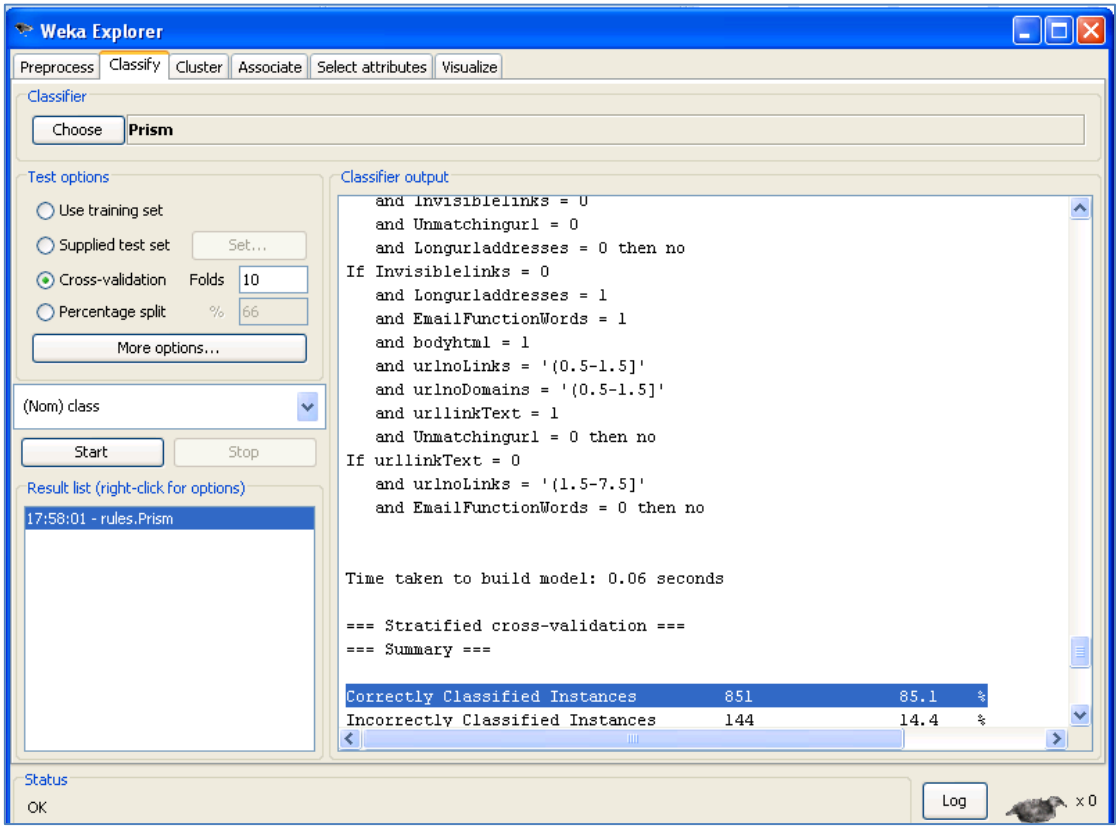

Figure 4.6The accuracy result of Prism algorithm (WEKA, 2002)

4.4 Precision and Recall Results

Precision and recall are two evaluation techniques are used in binary classification problem. They are calculated based on a matrix named confusion matrix as shown in Table 4.1. These two evaluation measures are computed as follows in equation (4.1) and (4.2):

$$\textbf{Precision} = \frac{TP}{TP+FP} \quad (4.1)$$

$$\textbf{Recall} = \frac{TP}{TP+FN} \quad (4.2)$$

Where:

True Positive (TP): represents the number of correct hits of positive instance.

False Negative (FN): gives the number of in-correct hits of positive instance.

False Positive (FP): denotes the number of in-correct hits of negative instance,

True Negative (TN): refers to the number of correct hits of negative instance.

Table 4.1: Confusion Matrix

	Classified Positive	Classified Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Figure 4.7 shows AveragePrecision and Recall among the proposed model and the other algorithms depending on the confusion matrix.

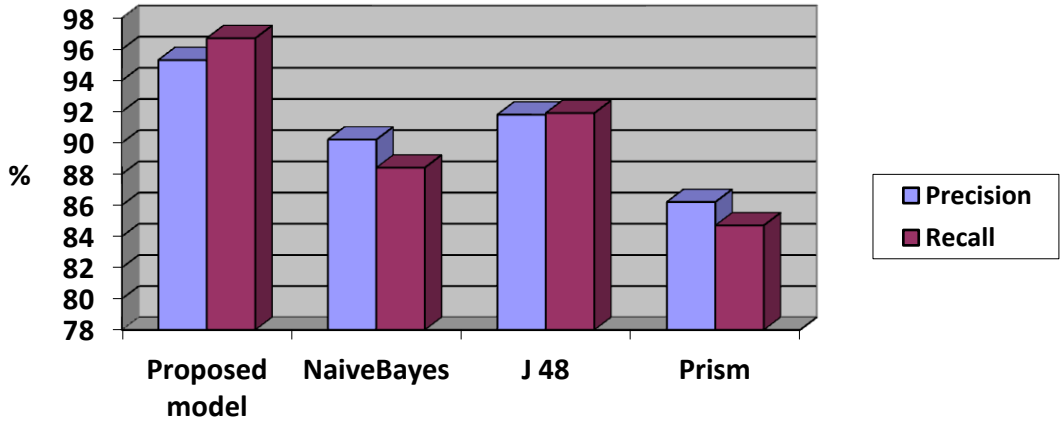

Figure 4.7 Average Precision and Recall results

The proposed model obtains the high average Precision and Recall results. This result means that an algorithm returned most of the relevant results when it gets the

high Recallresult. On the other hand, high Precision means that an algorithmreturned more relevant results than irrelevant among other classification algorithms which are used in the experiments, sinceit employees multiple rules for assigning the class for test data rather than using a single rule prediction.

4.5 Number of Rules

We have compared number of rules generated from our algorithm with Prism algorithmonly because it produces If-then rules just like AC does. As shown in Figure 4.8 we have produced less number of rules than were generated by Prism by using our new pruning procedure without effect accuracy rate. This procedure uses partly matching between rule body and training example which minimizes the size of the classifiers and reduces overfitting which led to have a less size classification system. To illustrate this procedure consider Table 4.2which handles example of training data to test the pruning step.

Table 4.2 Set of instances for testing pruning step

Body	Class
AB	L1
AB	L1
AB	L1
AB	L2
AB	L1
AB	L1
AB	L3
AB	L3

Now let test the rule "A & B → L1" over the training data in Table 4.2. The classifier building phase using the proposed new pruning technique will be removed all training instances in the table causing minimize the classifier size sinceit does not recognize the class in pruning step as well as using partial matching.

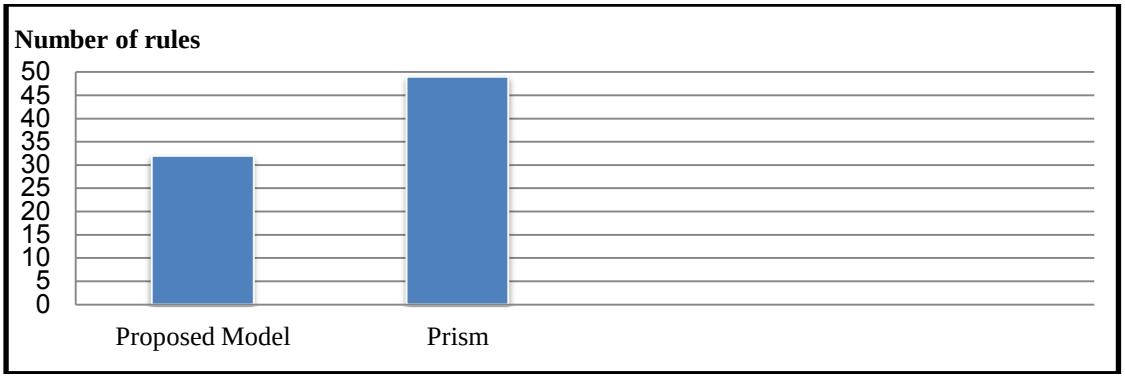

Figure 4.8: The number of rules generated

Chapter Five

Conclusion and Future Works

5.1 Conclusion

We have proposed a new algorithm using AC approach to handle the phishy emails problem. We have gatheredphishy and legitimate emails to train our model and test the applicability of AC in this kind of problem. This problem has many effects in our life causing different kind of losing. There are many exists solutions based on many sciences like statistical and probability, these solution did not have a 100% accuracy. However, we are achieved higher accuracy rate between different algorithms, NaiveBayes, J48 and Prism which reflect that the AC approach is effective to deal with such problem.

Moreover, the proposed algorithm is based on a new techniques were used in rules ranking, rule pruning and prediction step offer a good rule reductions let the classifier has a less number of rules where decrease redundant rule without effect the accuracy of the model. In particular, the ranking rule is used the superior rule on the top of the classifier, such rule has the maximum number of attributes of the antecedent side. This is since specific rules are more accurate in predicting test instance especially because they cover smaller number of training cases. Consequently, the algorithm applies partial matching as new criteria if full matching of the candidate rule body and the training set is not met in pruning step. This technique let the classifier contains less number of rules because a rule now has more training instance coverage.

The prediction phase in the proposed algorithm uses a new process to find all rules that match part of the test instance and calculates the average confidence of rules that have the same class then applies the class of the highest average value of confidence. So the prediction here is based on mathematical formula when the exact match fail in classify the test case. In addition, we have gathered phishy and legitimate emails with 23 features and offer feature assessment study over them.

5.2 FutureWorks

As a future work, the usage of genetic algorithm could offer extracting more features than the features are used now. Consequently, these features could discoverthe new phisher's tricks while this kind of problem always follows new different tricks to lure their victims. Also using fuzzy logic could offer nominal values for the numeric attributes which led to decrease the pre-processing time.Moreover, study a solution works with both phishy emails and phishy websites in the same time, by finding the set of features that could be used with both of them at the same time.

The thesis proposed new procedures in different steps: rule ranking, rule pruning and predicting the class for new case.The pruning step uses the partial matching without consider the class so if the same procedure applies with respect to the class attributes. Does it increase the classifier accuracy? And the number of pruned rules will decrease sharply. Moreover, investigate if partial matching with specific number of attributes effect the accuracy rate or matching in any attribute stays better as used in this thesis.

References

Abu-nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007).A Comparison of Machine Learning Techniques for Phishing Detection. SMU HACNet Lab Southern Methodist University Dallas,TX 75275

Abu-nimeh, S., Nappa, D., Wang, X., & Nair, S. (2009).Hardening Email Security via Bayesian Additive Regression Trees.SMU HACNet Lab, Southern Methodist University Dallas, TX, USA, (February).

Activity, P., & Report, T. (2012). Phishing Activity Trends Report 1 Quarter.

Adida, B., Chau, D., &Hohenberger, S. (2006). Lightweight Email Signatures. Computer Science and Artifical Intelligence Laboratory; Massachusetts Institute of Technology; 32 Vassar Street; Cambridge, MA 02139, USA 1-20.

Almomani, A., Wan, T., Altaher, A., Manasrah, A., Almomani, E., Anbar, M., Alomari, E., et al. (2012). Evolving Fuzzy Neural Network for Phishing Emails Detection National Advanced IPv6 Centre (NAV6), School of Computer Sciences , Faculty of Information Technology and Computer Sciences, 1099–1107.

Al-Momani, A., Ali, D., Tat-Chee, W., Al-Saedi, K., Altyeb, A., Sureswaran, R., et al., 2011.An Online Model on Evolving Phishing E-mail Detection and Classification Method.Journal of Applied Sciences, 3301-3307.

Alseadoon, I., Chan, T., Foo, E., & Nieto, J. (2012). Who is more susceptible to phishing emails. 23rd Australasian Conference on Information Systems 3-5 Dec 2012, Geelong , 1–11.

Amelia Z., Eva L., Gibaja, S. (2011). Multiple Instance Learning with Multiple Objective Genetic Programming for Web Mining. Elsevier Science Publishers B. V.Amsterdam, The Netherlands, 93–102.

Asanka, N., Arachchilage, G., Love, S., & Scott, M. (2012). Designing a Mobile Game to Teach Conceptual Knowledge of Avoiding " Phishing Attacks ." International Journal for e-Learning Security, 127–132.

Baralis, E., Chiusano, S., & Garza, P. (2008).A Lazy Approach to Associative Classification. IEEE Transactions on Knowledge and Data Engineering, 156–171.

Bjorn, B., Siegfried, N., and A. Z. (2011). Pattern-Based Classification: A Unifying Perspective. Department of Computer Science KatholiekeUniversiteit Leuven Celestijnenlaan 200A, 3001, Leuven, Belgium, 1–15.

Burges, C. (1997). A Tutorial on Support Vector Machines for Pattern Recognition.Kluwer Academic Publishers, Boston.Manufactured in the Netherlands.1-43.

Cao, Y., Han, W., & Le, Y. (2008). Anti-phishing based on automated individual white-list. Proceedings of the 4th ACM workshop on Digital identity management DIM '08, 51. New York, New York, USA: ACM Press.

Chen, M., Huang, C., Chen, K., & Wu, H. (2005). Aggregation of orders in distribution centers using data mining. ELSEVIER, 28(3), 453–460.

Damodaram, R., Phil, M., &Valarmathi, M. L. (2012). Phishing website detection and optimization using Modified bat algorithm. International Journal of Engineering Research and Applications, 870–876.

Data Protection & Breach Readiness Guide. (2012). 2012 Data Protection & Breach Readiness Guide. © 2012 Online Trust Alliance (OTA), 1–28.

European, I., Data, C., & Science, C. (2007). Data mining techniques for suspicious email detection: A comparative study. IADIS European Conference Data Ming 2007.213–217.

Fayyad, U., and K. I. (1993). Multi-interval discritisation of continues-valued attributes for classification learning, California institute technology, USA 1022-1027.

Gambhir, S., &Gondaliya, P. (2012).Misclassification Penalties in Associative Classification. International Journal of Engineering Research & Technology (IJERT), 1–6.

Guofei G., Phillip P., Vinod Y., Martin F., W. L. (2007). Detecting Malware Infection Through IDS-Driven Dialog Correlation. 16th USENIX Security Symposium, 35.

Gupta, S., &Todwal, V. (2012). Web Data Mining & Applications. International Journal of Engineering and Advanced Technology (IJEAT), 20–24.

Habib, A., Hoque, A., &Rahman, M. (2012).High Performance Query Operations on Compressed Database. International Journal of Database Theory and Application Vol. 5, No. 3, September, 2012, 5(3), 1–14.

Irani, D., Webb, S., Giffin, J., &Pu, C. (2008). Evolutionary Study of Phishing.College of Computing Georgia Institute of Technology Atlanta, Georgia.

Islam, M. R., Abawajy, J., & Warren, M. (2009). Multi-tier Phishing Email Classification with an Impact of Classifier Rescheduling. 2009 10th International Symposium on Pervasive Systems, Algorithms, and Networks, 789–793.

Karthikeyan, M., Suriya K. (2012).Literature review on the data mining and information security. International Jornal of Computer Engineering and Technelogy (IJCET), 141–146.

Kenkel, B., and Curtis S. (2011).Data Mining for Theorists.Department of Political Science, University of Rochester 1–26.

Khonji, M., & Iraqi, Y. (2011).Lexical URL analysis for discriminating phishing and legitimate e-mail messages. Internet Technology and Secured, 11–14.

Kingdom, U., States, U., & Africa, S. (2012). Global fraud losses down despite a 19 percent increase in phishy attacks. EMC Corporation. EMC, RSA.

Kovacs, E. (2012) http://news.softpedia.com/news/RSA-Phishing-Attacks-Worldwide-Cause-Losses-of-687M-556M-in-H1-2012-287534.shtml.

Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., & Hong, J. (2010).Teaching Johnny not to fall for phish.ACM Transactions on Internet Technology.

Kundu, G., Munir, S., & Bari, F. (2009).A Novel Algorithm for Associative Classification. Knowledge Acquisition: Approaches, Algorithms and Applications Springer-Verlag Berlin, Heidelberg, 61-75.

Li, W., Han, J., & Pei, J. (2001). CMAR: Accurate and efficient classification based on multiple class-association rules. IEEE International Conference on Data Mining, ICDM.

Li, X., & Yu, D. (2008). ACCF: Associative Classification Based on Closed Frequent Itemsets. Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on: Oct. 2008, 380 – 384.

Lim, T., Loh, W. and Shih, Y. (2000).A comparison of prediction accuracy, complexity and training time of thirty-three old and new classification algorithms.203-228.

Liu, B., Hsu, W. and Ma, Y. (1998).Integrating classification and association rule mining. Proceedings of the KDD New York, NY, 80-86.

Ma, L., Ofoghi, B., Watters, P., & Brown, S. (2009). Detecting Phishing Emails Using Hybrid Features. 2009 Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing IEEE, 493-497.

Qazafi, M., Thabtah, F., and McCluskey, T. (2007). Looking at the class associative classification.School of Computing and Engineering Researchers' Conference, University of Huddersfield.

Marko, T. (2012): http://slavnik.fe.uni-lj.si/markot/csv2arff/csv2arff.php.

Martino, A., &Perramon, X. (2010). Phishing Secrets: History , Effects , and Countermeasures. International Journal of Network Security, 163–171.

Millersmiles (2012):http://www.millersmiles.co.uk/.

Mishra, P., Padhy, N., &Panigrahi, R. (2012).The survey of data mining application and feature scope. Asian Journal of Computer Science and Information Technology, 68–77.

Nikolov, P. (2011). Aggregate Queries in NoSQL Cloud Data Stores Master's Thesis, PDCS Submitted to the Department of Sciences, VrijeUniversiteit, Amsterdam, The Netherlands.

Olaru, C., &Wehenkel, L. (2003).A complete fuzzy decision tree technique. Fuzzy Sets and Systems, 2003 Elsevier B.V. 221–254.

Olivo C., Santin A., & Oliveira, L. (2011). Obtaining the threat model for e-mail phishing. ELSEVIER, 4–11.

Paaß, G., &Bergholz, A. (2009).AntiPhish - Machine Learning for Phishing Detection., Project Exhibition at ECML/PKDD 2009.7-8.

Paaß, G., Reichartz, F., &Strobel, S. (2008).Improved Phishing Detection using Model-Based Features.Fraunhofer IAIS SchloßBirlinghoven 53754 St. Augustin, Germany.

Parmar, B. (2012). Protecting against spear-phishing. Computer Fraud & Security January 2012.

Quinlan, J. and Cameron-Jones, R. (1993). FOIL: A midterm report. Proceedings of the European Conference on Machine Learning, Vienna, Austria.3-20.

Ramanathan, V., & Wechsler, H. (2012).phishGILLNET phishing detection using probabilistic latent semantic analysis. EURASIP Journal on Information Security, 2012(1).

Razvan, R., & Maria, M. (2010).The security of Web 2.0 and digital economy, 168–170.

Romero, C., & Ventura, S. (2010). Predicting School Failure Using Data Mining. University of Cordoba, Spain.

Saad, Y., Gao, D., Ngo, T., Bobbitt, S., Chelikowsky, J. R., &Andreoni, W. (2011). Data mining for materials: Computational experiments with AB compounds. University of Texas, Austin, 1–24.

Safavian, S. R., &Landgrebe, D. (1991).A Survey of Decision Tree Classifier Methodology. School of Electrical Engineering, 660–674.

Salama, M., Panda, M., Elbarawy, Y., Hassanien, A., & Abraham, A. (2012). Computational Social Networks: Security and Privacy, Springer-Verlag London.

Scamdex (2012): http ://www.scamdex.com/Phishing-index.php.

Shalendra, C. (2005). Fighting Spam, Phishing and Email Fraud.University of California.

Sheng, S., Wardman, B., Warner, G., Cranor, L., & Hong, J. (2009).An Empirical Analysis of Phishing Blacklists.Sixth Conference on Email and AntiSpam July 1617, 2009, Mountain View, California USA.

Shih, C., &Kochanski, G. (2006). Bayes Theorem ,This work is available under http://kochanski.org/gpk/teaching/0601Oxford.1-10.

Soni, S., &Vyas, O. (2010).Using Associative Classifiers for Predictive Analysis in Health Care Data Mining.International Journal of Computer Applications.

SonicWall. (2008). Bayesian Spam Classification Applied to Phishing E-Mail. 2008 SonicWall.

Spam-assassin (2012):http://csmining.org/index.php/spam-assassin-datasets.html.

Spywareremove (2013): http://www.spywareremove.com/removewindowshealthkeeper.html.

Tang, Z., & Liao, Q. (2007). A New Class Based Associative Classification Algorithm. IAENG International Journal of Applied Mathematics, 1–5.

Thabtah, F., Hadi, W., Abdel-jaber H., Abdelhamid N. (2010).New Rule Pruning Methods in Associative Learning Text Mining.Journal of Cognitive Computing.Springer-verlag.

Thabtah, F., & Cowling, P. (2005). MCAR: multi-class classification based on association rule. The 3rd ACS/IEEE International Conference on Computer Systems and Applications, 2005., 130–136.

Thabtah, F., & Cowling, P. (2008).Mining the data from a hyperheuristic approach using associative classification. Expert Systems with Applications 2007 Elsevier Ltd, 1093–1101.

Thareja, P., Goyal, A. (2011). The Edge of Character in Know-All Learning.The Journal of Consultancy Development Centre.

Toolan, F., &Carthy, J. (2010). Feature Selection for Spam and Phishing Detection. Group.2010 IEEE.

WEKA (2002): Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka.

Yearwood, J., Mammadov, M., & Banerjee, A. (2010). Profiling Phishing Emails Based on Hyperlink Information. 2010 International Conference on Advances in Social Networks Analysis and Mining, 120–127.

Yin X. and Han J. (2003). CPAR: Classification based on predictive association rule. Proceedings of the SDM (pp. 369-376). San Francisco, CA.

Zaki, M. (2002). CHARM: An Efficient Algorithm for Closed Itemset Mining. Computer Science Department, Rensselaer Polytechnic Institute.

Zaki, M., &Parthasarathy, S. (1997).New Algorithms for Fast Discovery of Association Rules.The University of Rochester Computer Science Department & NSF Research Initiation Award.

Zhu, Y., Luo, W., Chen, G., &Ou, J. (2012). A Multi-label Classification Method Based on Associative. Journal of Computational Information Systems, 791–799.

ملخص

يعتبر البريد إلكتروني احد افضل طرق الإتصال في الوقت الحاضر. في اليوم الواحد يُرسل الناس ويَستلمونَ العديد مِنْ الرسائل، للاتصال مَع بعضهم البعض ويُتبادلونَ الملفاتَ والمعلوماتَ. لكن يعتبر البريد الإلكتروني الخادع الجريمة الأكثر شيوعاً مِنْ جرائم الانترنت الإلكترونيةِ في الوقت الحاضر. وهذه المشكله تعتبر أحد مشاكل تقنياتِ الهندسة الإجتماعيةِ حيث أن جهل المستخدم يَسْمحُ للناس السيئين لاستغلال الضعفِ في تقنياتِ أمن الانترنت. الهدف هو مُحاولُة الحُصُول على معلوماتِ سرّيةِ، مثل أسماءِ المستعملين، كلمات سر، أوراق إعتماد حسابِ ماليةِ وتفاصيل بطاقةِ إئتمان.

هدف الرسالة هو تصنيف الايميل باستخدام ِالتصنيفِ الِترابُطي. حيث تَتحرّى هذه الإطروحةِ تطبيقَ التصنيفِ الِترابُطي في المشكلةِ المعقدةِ للايميل الخادع لإثتاج قواعد شرطيه عن طرق تقديم طرق جديدة في مرحلة ترتيب القواعد الشرطيه. وتم ايضا طرح طريقة جديدة في عمليه تقليل عدد القواعد وكذلك تم استخدام معادلة رياضية في عملية التنبؤ عن الايميل غير المصنف.

أيضاً الدراسة وفرت تقييم عن مجموعة من الميزّاتَ الأكثر تكراراً التي يُمْكِنُ أنْ تُصنّفَ البريد إلكتروني إلى الصنفِ الصحيح. يُعالجُ قسمُ التقييمَ تجاربَ شاملة على البيانات التي تَستعملهاُ الخوارزمية المُقتَرَحة وخوارزميات التصنيفِ المعروفةِ الأخرى عن طريق المقارنه بينهم عن طريق نسبةِ الدقةِ حيث ان الخوارزمية المُقتَرَحة في هذا العمل حَصلتْ على أعلى نسبه من الدقيقةِ. وايضا عن طريق انتاج عدد اقل من القواعد الشرطيه من الخوارزميات الاخرى.

أستخدام تقنيات التنقيب عن البيانات لاكتشاف الايميل الخادع

من قبل

بدر عبدالحفيظ أحمد الرحامنه

بإشراف

د.فادي فايز

قدمت هذه الرسالة إستكمالا لمتطلبات الحصول على درجة الماجستير في علم الحاسوب

عمادة البحث العلمي والدرسات العليا

جامعة فيلادلفيا

كانون ثاني 2013