ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A LITERATURE REVIEW ON PHISHING EMAIL DETECTION USING DATA MINING

Jagruti Patel*, Sheetal Mehta
Information Technology Department, Parul Institute of Engineering and Technology,Limda, Vadodara-390001, India,
Computer Science and Engineering Department, Parul Institute of Engineering and Technology,*Limda, Vadodara-390001, India*

ABSTRACT

Fraud emails have become common problem in recent years. Fraud emails are a real threat to internet communication. In this paper, hybrid features are used for detecting fraud emails to determine how fast they have classified fraud emails and normal emails. Instead of hybrid feature, only content as a feature can also be used but most of the phishing email has similar content as normal email, so detection of phishing email is more complex and these approaches cannot give the higher rate classification. Hybrid feature selection approach based on combination of content based and header information. It presents an overview of the various techniques presently used to detect phishing email.

KEYWORDS: Mining, Classification, Internet security

Introduction

Now a days phishing attack growing significantly in each year. In phishing attack, attacker sends a mail that has a link in which when user clicks on that link, the users have to fill information like account details, password etc. on that page. After user fills the information it directly can be accessed by the attacker and there will be misuse of the private information.
There are different types of fraud email: [1]
1. Spam email
The Spam emails are sent for different intensions, mainly by advertisement popup. The Spam emails are usually sent into bulk. These emails are not as much harmful as phishing emails are. That type of mail cause the more CPU usage time and wastes the resources.
2. Suspicious email
Suspicious emails are another category of fraud email. Suspicious emails are those which contain some materials which are worth analysis. Suspicious email may contain some clues regarding some malicious activities.
3. Phishing email
The detection of phishing email is hard problem because phishing emails are normally same as legitimate emails. In phishing email, Phisher (attacker) is sending an email which contain some link when user click on that URL (phishing link) the phishing web page looks like a legitimate web page, in which user have to fill information like personal information, account information or password, etc. Among the users few users clicks on that URL (phishing link) which is embedded on that email which is sent by the "phisher". When user fills that information it redirects to the attacker and attacker can use these information.
There are different types of phishing email :

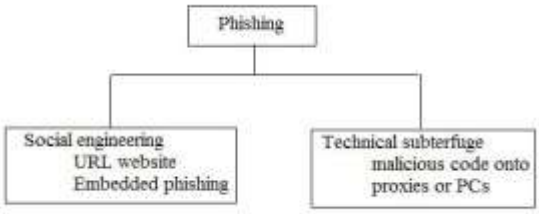

Fig 1 Trusted Bank phishing email [9]

In the first technique, social engineering schemes, It depends on forged email claims that appear to originate from a legitimate company or bank. Through an embedded link within the email, the phisher attempt to redirect users to fake websites. These fake websites are used for obtaining financial data(user names, password, credit card numbers and personal information) from victims.
The second technique involves technical schemes that are malicious code or malicious link embedded in the email, or by detecting and using security holes

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

in the user's computer to obtain the victim's online account information directly.

MATERIALS AND METHODS

Phishing Email Detection Techniques

Subheading should be 10pt Times new Roman, justified.

Many approaches against phishing attacks have been proposed in the literature. These protection approaches against phishing attacks are discussed below:

A. Network level protection

Network level protection based approach is used for allowing a website administrator to block messages that usually send fraud emails. Some websites are defined as a blacklist. So when these types of website are detected, tool that is describe below gives pop-up alert. An attacker or phisher can avoid this protection technique by controlling legitimate user's PCs or by continuously changing IP addresses.

Snort is open source software also employed at the network level. Rules in Snort are constantly updated to maintain protection. [19]

Comparison of the two phishing attack detection tools at the network level is presented in Table 1

Tool	Description	Advantages	Disadvantages
Domain name system blacklists	Database used by internet service providers	An updates list of offending addresses	Phisher can easily evade
Snort	Heuristic/rule engine	Good at detecting level attacks	-Rule require manual adjustments - Does not look at content of message

Table 1 Tool used at network level phishing email protection

B. Authentication

Authentication based approaches for confirming that the email sent is from valid path and domain name is not form blacklist / not spoofed by phisher.

Email authentication is done by sending the hash of the password with the domain name using digital signature and password hashing. PGP and S/MIME are examples of digital signature technologies.

C. Client side tools

Tools that used on the client side include user profile filters and browser-based toolbars.

Spoof Guard, Net Craft[16], Calling ID[17], Cloud Mark[18], eBay toolbar and IE phishing filter are some of the client side tools. They include a study of phishing and attack by detecting phishing "Web browsers" directly.

Client side tools which is used for domain checks, URL check, input, page content and algorithms. These tools, which are designed and trained, using typical prototypes of phishing website URLs, a dialog box is used for warning.

These tools also depend on black- listing and white-listing, which is a technique used to prevent phishing attacks by checking URL embedded in emails or by checking the website directly.

In the Mozilla Firefox browser, each Web page selected by a user is tested against a blacklist of well-known phishing Websites.

In the black-listing process, a list of the detected phishing Websites is automatically downloaded to the user machine with updates at standard intervals. The average threat time of an online phishing Website is three days and sometimes the sites are blacklisted within a few hours. However, this technique does require time for a new phishing Website to be reported and added to the blacklist. Blacklisting can also produce false negatives and miss many phishing emails; therefore, it is not particularly effective. Blacklists are ineffective in protecting users from 'fresh' phishing email, as most of them blocked less than 20% of phish at hour zero.

White-listing is a collection of "good" URL compared to outside links in receiving incoming emails. It appears more promising, however, producing a list of trustworthy sources is time-consuming, and it is a huge task. Two problems encountered by this technique are its producing a high number of false positives, allowing phish to get through, and its filtering of ham emails. Therefore, white-listing is not effective enough to be used for detecting phishing attacks.

Black-listing and white-listing techniques are very weak to work with technology changes (like IPV4 versus IPV6, tiny URLs, etc.). Moreover, most of users do not give attention to the warning dialogs. Due to above mentioned weaknesses; these techniques are not an effective solution to detect zero day attack.

D. User education

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

User education, based on social response approaches, depend on increasing the level of awareness and education about phishing attacks.
Approach offers online information about the risks of phishing attacks, and how to keep away from this attack. These materials are frequently published by the governments, non-profit organizations from trading platforms, such as eBay, Amazon, and Bank of America to financial enterprise.
After such training, users can able to detect phishing email. Training system provides a warning along with active items using text and graphics.
E. Server side filters and classifiers
Server side filters, based on content-based filtering approaches, are considered as the best option for zero day attacks. Therefore, most researchers try to solve zero day attack from this side. This depends on an extracted set of phishing email features. These features are trained on machine-learning algorithms by classifier to classify as ham (legitimate) email or phishing email. After that, this classifier may be used on an email to predict the class of newly received emails. For their product advertisement, Spammeruse the internet and sends a spam mail to the huge number of user. Phishing emails are normally looking like a normal mail and look like that it comes from the trustworthy companies. Therefore, many techniques used in spam detection cannot be used in phishing email detection.

2.1 Feature selection for detecting phishing email:
The below figure shows that normally email contain two parts:
Header: It contains sender and receiver information
Body: It contains content of the email

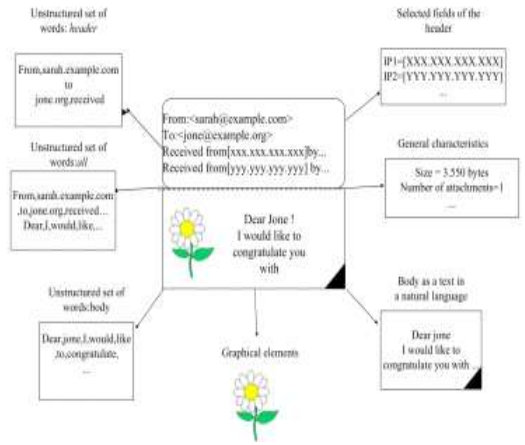

Fig 3 Email structure for feature selection [13]

Feature selection set is classified into three sets for detecting phishing email. That is given below [9]:

1) Basic Features: That is directly extracted from email without any processing. It also categorized into different sets, That are as follows
 Structural features: Extracted from an HTML tree, structure of email body.
 Link features: Different features of URL links embedded in an email, Number of link with IP, Number of URL visible to user, number of links behind an image, number of dots in a link and so on.
 Element features: Type of web technology used in email such as HTML, scripting (JavaScript and any other)
 Word list feature: List of words are used for detecting phishing email and classified by Boolean features, whether it occurs in email or not. Word stems such as account, update, confirm, verify, log, clicks and so on.
2) Latent topic model features: Cluster of words that appear together in email. The words "click" and "account" often appear together. Classified based on different categories like financial, family etc.

Authors	Number of feature s	Feature Approach	Sample	Accuracy
Fette et al	10	URL and script Based	Phishing - 860 Non phishing - 6950	97.6% F-measure and false positive rate of 0.13% and a false negative rate of 3.6%.
Abu-Nihmeh et al	43	Keyword Based	1700 legitimate emails and 1700 phishing emails	F-measure of 90%.
Basnet et al	16	URL and content based	4000 emails legitimate and 973 phishing emails	highest accuracy of 97.99% with BSVM and NN.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Toolan et al.	22	Behavioral based	Total dataset 6097 Non-phishing 70% and spam 30%	dataset1: 97% dataset2: 84% dataset3: 79% f1-score of 99.31%
Ammar et al.	16	HTML part and URL based	1000 Datasets	error rates of 0.13and 0.12
Bergholz et al	27	Model based	5000 Datasets	f1-score of 99.46%.
Mayankpandey et al.	23	Content based	2500 phishing and non-phishing emails.	classifers GP achieved good accuracy with feature selection and without feature selection compare to others
Isredza Rahmi et al.	7	Hybrid features	6923 datasets	96% accuarcy and 4% False positive and False Negative rate.
Mingxing He et al	12	Content and URL based	Dataset1 :100 login pages Dataset2:100 phishing web pages and 100 normal web pages	high detection rate and low false positive rate
Sarwat Nizamani et al		Content based	Total 8000 emails 2500 emails are fraudulent	Accuracy 96%

3. Phishing email evaluation methods:[10]

1. True Positive (TP): The number of phishing email correctly classified as phishing:
TP= np --> p/ Np
2. True Negative (TN): The number of ham emails correctly classified as phishing:
TN=nh --> h / Nh
3. False positive (FP): The number of ham email wrongly classified as phishing:
FP=nh --> p / Nh
4. False Negative (FN): The number of phishing emails wrongly classified as ham:
FN=nh --> h / Np
5. Precision (p): Measures the rate of correctly detected phishing attacks in relation to all instances that were detected as phishing
p= |TP| / (|TP|+|FP|)
6. Recall(r): Measures the rate of correctly detected phishing attacks in relation to all existing phishing attacks.
r= |TP| / (|TP|+|FN|)
7. F1 score: Harmonic mean of P and R.
f1= (2p.r) / (p+r)
8. Accuracy: The percentage of correct prediction [12]
Accuracy = (|TP| + |TN|) / (|TP| + |TN| + |FP| + |FN|)

whereNh = total number of ham emails
nh-->h = number of ham emails classified as ham
nh-->p = number of ham emails misclassified as phishing
Np = total number of phishing emails
np-->h = number of phishing email misclassified as ham
np-->p number of phishing emails classified as phishing emails

3.1 Techniques used

1) Methods based on Bag-of-Words model: This method is a phishing email filter that considers the input data to be a formless set of words that can be implemented either on a portion or on the entire email message. It is based in machine learning classifier algorithms.

Some classifiers and approaches related to this method appear below.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

A. Support vector machine (SVM) [12]: One of the most commonly used classifier in phishing email detection. In 2006, the SVM classifier was proposed for phishing email filtering. SVM worked based on training email samples and a pre-defined transformation θ: Rs→ F, which builds a map from features to produce a transformed feature space, storing the email samples of the two classes with a hyper plane in the transformed feature space shown in Figure 4.

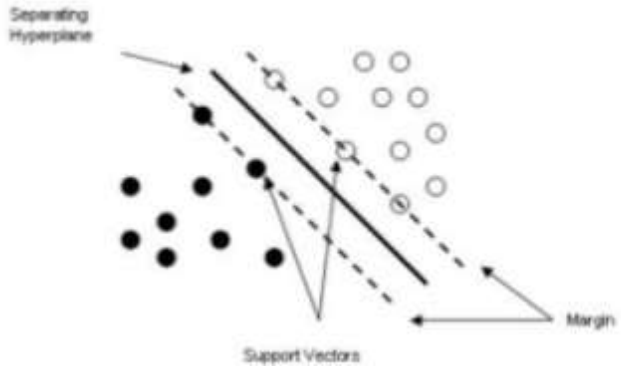

Fig: 4 Support Vector Machine

B. K-NearestNeighbor (k-NN): Classifier proposed for phishing email filtering. Using this classifier, the decision is made as follows: based on k-nearest training input, samples are chosen using a pre-defined similarity function; after that, the email x is labeled as belonging to the same class as the bulk among this set of k sample (Figure 5)

C. Naive bays classifier: Simple probabilistic classifier, which works based on Bayes' theorem with powerful "naive" independence assumptions Ganger [9]. This classifier, used in text classification, can be a learning-based variant of keyword filtering. To ensure preciseness, all features are statistically independent.

D. Boosting: a boosting algorithm combines many hypotheses like "One-level decision trees." The main idea of this algorithm depends on sequential adjustments at each phase of the classification process where a fragile (not very accurate) learner is trained. The output results of each phase are used to reweigh the data for future stages. The larger weight is assigned to the input samples that are misclassified. Term frequency-inverse document frequency (TF-IDF) is use for word weights, as features for the clustering. The document frequency of the word w is implemented by DF(w) which is defined as the number of email messages in the collected data set where the word w appears in the document at least once as shown in the formula [20].

$Wxy = TFxy \cdot \log x$

Where Wxy is the weight of xth Word in the yth document (email), TFxy is the occurrences number of the xth word (w) in the yth document (email), DFx is the number of email messages in which the ith word (w) occurs, and S, as above, is the total number of messages in the training dataset.

Bag-of-Words model has many limitations. It is implemented with a large number of features, consumes memory and time, and mostly works with a supervised learning algorithm. Furthermore, it is not effective with zero day attack.

2)Multi Classifiers Algorithms [10]: These approaches in general depend on comparison between sets of classifiers.

Presently, more and more research has used new classifier algorithms like Random Forests (RF). RFs are classifiers which merge several tree predictors, where each tree depends on the values of a random Vector sampled separately, and can handle large numbers of variables in a data set.

Another algorithm, Logistic Regression (LR), is one of the most widely used statistical models in several fields for binary data prediction. It used because of its simplicity.

Neural Networks (NNet) classifiers, which consist of three layers (input layer, hidden layer, and output layer), gains the requisite knowledge by training the system with both the input and output of the preferred problem. The network is refined until results have reached acceptable accuracy levels as shown in Figure 6.

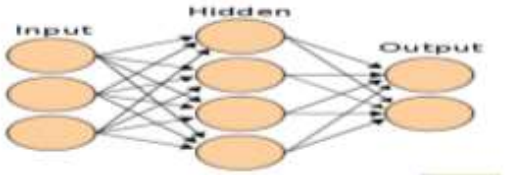

Fig:6 Neural network

The power of NNet comes from the nonlinearity of the hidden neuron layers. Nonlinearity is important for the network learning of complex mappings. Sigmoid function is the commonly-used function in neural networks. Abu-Nimeh et al. [10] compared six classifiers relating to machine learning technique for phishing prediction, namely, Bayesian Additive Regression Trees (BART), LR SVM, RF, NNet, and Classification and Regression Trees (CART).

3) Classifiers model based features: These approaches build full models that are able to create

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

new features with many adaptive algorithms and classifiers to produce the final results.

4) Clustering of phishing email: Clustering is the process of defining data grouped together according to similarity. It is usually an unsupervised machine learning algorithm.

Technique used	Advantages	Disadvantages
Methods based on Bag –of –words model	-Build good scanner between user's mail transfer Agent(MTA) and mail user Agent(MUA)	-Huge no. of features consumes memory -mostly working with supervised learning algorithm -fixed rules -weak detection of zero-day attack
Multi classifiers algorithms	-provide clear idea about the effective level of each classifier on phishing email	-Nonstandard classifier -Mostly working with supervised learning -weak in zero day detection
Classifier Model based features	-High level of accuracy -create new type of feature like Markov features	-huge number of feature -time consuming -Higher cost
Clustering of phishing email	-Fast in classification process	-Less accuracy
Multi-layer system	-High level accuracy	-Time consuming

CONCLUSION

Phishing emails have become common problem in recent years. Phishing is a type of attack in which victims sent emails into which users have to provide sensitive information and then it directly sent to the phisher. So detection of that type of email is necessary. There are many techniques for detecting phishing email but there is some limitation like accuracy is low, content can be same as legitimate email so cannot be detected, detection rate is not high. So some advance method is required. To overcome that limitation, hybrid feature selection can be applied. The features are based header information and URL. By using header information, sender's behavior can be analyzed. By applying this approach in future, the accuracy and detection rate can be measured.

ACKNOWLEDGEMENTS.

This work is supported and guided by my research guide. I am very thankful to my research guide Mrs. Sheetal Mehta, Assistant Professor, CSE Department, Parul Institute Of Engineering And Technology, Gujarat, India for supporting me.

REFERENCES

[1] Xue Li, , Vasu D. Chakravarthy, , Bin Wang, and Zhiqiang Wu, "Spreading Code Design of Adaptive Non-Contiguous SOFDM for Dynamic Spectrum Access" in IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 5, NO. 1, FEBRUARY 2011

[2] J. D. Poston and W. D. Horne, "Discontiguous OFDM considerations for dynamic spectrum access in idel TV channels," in Proc. IEEE DySPAN, 2005.

[3] R. Rajbanshi, Q. Chen, A.Wyglinski, G. Minden, and J. Evans, "Quantitative comparison of agile modulation technique for cognitive radio tranceivers," in Proc. IEEE CCNC, Jan. 2007, pp. 1144–1148.

[4] V. Chakravarthy, X. Li, Z. Wu, M. Temple, and F. Garber, "Novel overlay/underlay cognitive radio waveforms using SD-SMSE framework to enhance spectrum efficiency—Part I," IEEE Trans. Commun., vol. 57, no. 12, pp. 3794–3804, Dec. 2009.

[5] V. Chakravarthy, Z. Wu, A. Shaw, M. Temple, R. Kannan, and F. Garber, "A general overlay/underlay analytic expression for cognitive radio waveforms," in Proc. Int. Waveform Diversity Design Conf., 2007.

[6] V. Chakravarthy, Z. Wu, M. Temple, F. Garber, and X. Li, "Cognitive radio centric overlay-underlay waveform," in Proc. 3rd IEEE Symp. New Frontiers Dynamic Spectrum Access Netw., 2008, pp. 1–10.

[7] X. Li, R. Zhou, V. Chakravarthy, and Z. Wu, "Intercarrier interference immune single carrier OFDM via magnitude shift keying modulation," in Proc. IEEE Global Telecomm. Conf. GLOBECOM , Dec. 2009, pp. 1–6.

[8] Parsaee, G.; Yarali, A., "OFDMA for the 4th generation cellular networks" in Proc. IEEE Electrical and Computer Engineering, Vol.4, pp. 2325 - 2330, May 2004.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

[9] 3GPP R1-050971,"R1-050971 Single Carrier Uplink Options for EUTRA: IFDMA/DFT-SOFDM Discussion and Initial Performance Results ",http://www.3GPP.org,Aug 2005
[10] IEEE P802.16e/D12,'Draft IEEE Standard for Local and metropolitan area networks--Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems', October 2005
[11] 3GPP RP-040461, Study Item: Evolved UTRA and UTRAN, December 200
[12] R. Mirghani, and M. Ghavami, "Comparison between Wavelet-based and Fourier-based Multicarrier UWB Systems", IET Communications, Vol. 2, Issue 2, pp. 353-358, 2008.
[13] R. Dilmirghani, M. Ghavami, "Wavelet Vs Fourier Based UWB Systems", 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, pp.1-5, Sep. 2007.
[14] M. Weeks, Digital Signal Processing Using Matlab and Wavelets, Infinity Science Press LLC, 2007.
[15] S. R. Baig, F. U. Rehman, and M. J. Mughal, "Performance Comparison of DFT, Discrete Wavelet Packet and Wavelet Transforms in an OFDM Transceiver for Multipath Fading Channel,",9th IEEE International Multitopic Conference, pp. 1-6, Dec. 2005.
[16] N. Ahmed, Joint Detection Strategies for Orthogonal Frequency Division Multiplexing, Dissertation for Master of Science, Rice University, Houston, Texas. pp. 1-51, Apr. 2000.
[17] Sarwat Nizamani a,b,*, Nasrullah Memon a,c, Mathies Glasdam "Detection of fraudulent emails by employing advanced feature abundance"Elsevier 2014
[18] Mayank Pandey , Vadlamani Ravi* Institute for Development & Research in Banking Technology, Masab Tank, Hyderabad, India "Detecting phishing e-mails using Text and Data mining"IEEE2012
[19] Liping Ma,Bahadorrezda Ofoghi, Paul Watters, Simon Brown"Detecting Phishing Emails Using Hybrid Features"IEEE 2009
[20] Mingxing He a, Shi-Jinn Horng a,b,c,⇑, Pingzhi Fan c, MuhammaKhurram Khan d, Ray-Shine Run e,Jui-Lin Lai e, Rong-Jian Chen e, Adi Sutanto b"An efficient phishing webpage detector"Elsevier2011
[21] Toolan and J. Carthy, "Phishing detection using classifier ensembles," in eCrime Researchers Summit, IEEE Conf, Tacoma, WA, USA, 2009, pp. 1-9.
[22] Gansterer, W.N., Polz, D.: E-Mail Classification for Phishing Defense. LNCS Advances (2009)
[23] Mahmoud Khonji, Youssef Iraqi "Enhancing phishing email classifiers: A lexical URL Analysis Apporach" IJISR 2012
[24] Madhusudhanan Chandrasekaran, Krishnan Narayanan and Shambhu Upadhyaya "phishing email detection based on structural properties"
[25] Ammar Almomani, B. B. Gupta, Samer Atawneh, A. Meulenberg, and Eman Almomani" A Survey of Phishing Email Filtering Techniques" IEEE 2013
[26] S. Abu-Nimeh, et al., "A comparison of machine learning techniques for phishing detection," in Proc. eCrime Researchers Summit,, Pitts- burgh, ACM Conf, Pittsburgh, PA,2007, pp. 60-69.
[27] Andre Bergholz, Gerhard Paab, Jeong-Ho Chang"Improved phishing detection using model based feature" Springer ss2008
[28] Chai-mei-chen,D.J.Guen.Quen-kei-su,"Feature set identification for detecting suspicious URLs using Bayesian classification in social network" Elsevier 2014
Websites:
[29] Email structure:http://www.google.co.in/imgres
[30] Nazario:J. Phishing Corpus, http://www.monkey.org/jose/wiki/doku.php?id=phishingcorpus
[31] Spamassassin public corpus, http://spamassassin.apache.org/publiccorpus
[32] Netcraft."Netcraft toolbar",2006,Available: http://toolbar.netcraft.com/
[33] CallingID. "Your Protection from Identity Theft, Fraud, Scams and Malware",accessed 29 May 2012,Avail-able:http://www.callingid.com/Default.aspx
[34] CloudMark, accessed 29 may 2012, available: http://www.cloudmark.com/en/products/cloudmark-desktopone/index

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

[35] Snort Home Page. Accessed 29 May 2012, available: http://www.snort.org/
Books:
[36] Data Mining:Concepts and Techniques, Second Edition Jiawei Han-University of Illinois at Urbana-Champaign,Micheline Kamber