Comparison of the Intelligent Techniques for Data Mining in Spam Detection to Computer Networks

Kelton Costa, Patricia Ribeiro, Atair Camargo, Victor Rossi, Henrique Martins, Miguel Neves, Ricardo Fabris[1]
João Paulo Papa[2]
[1]College of Technology of São Paulo state
Bauru, Brazil
[2] Department of Computing - UNESP - University Paulista State
Bauru, Brazil
{kelton.costa, vic.v.rossi, henmartins, ricardoffabris}@gmail.com, patriciabellin@yahoo.com.br, junior.camargo@hotmail.co.uk, miguel.neves01@fatec.sp.gov.br, papa@fc.unesp.br

ABSTRACT: *Anomalies in computer networks has increased in the last decades and raised concern to create techniques to identify the unusual traffic patterns. This research aims to use data mining techniques in order to correctly identify these anomalies. Weka is a collection of machine learning algorithms for data mining tasks which was used to identify and analyze the anomalies of a data set called SPAMBASE in order to improve this environment.*

Keywords: Anomalies, Computer Networks, Data Mining, Artificial Neural Networks, Weka Tool, SPAMBASE

Received: 17 October 2013, Revised 2 December 2013, Accepted 8 December 2013

1. Introduction

Spam email damaging effects are currently investigated widely and the literature has extensive studies on it. As the spammers indulge in breaking the filters using sophisticated techniques, the researchers also generate newer measures to counter the attacks. The efforts include many tools and techniques and one significant among them is the use of artificial neural networks and data mining processes with decision trees in order to minimize the spam e-mails' damaging effects. Data Mining as a domain experiences new focus on varied applications across disciplines. It has potential in the spam detection.

Data mining is a part of KDD (Knowledge Discovery in Data bases) process which aims to select techniques to be used to find patterns in these data set in order to find correlated patterns about a specific interest [1] [2].

The steps for knowledge discovery in KDD [1] [2] can be presented in a cognitive, interactive and exploratory way based in the following stages: defining the type of knowledge to search for, defining a group or subgroup of data to search for, pre-processing, reduction of the data set, data mining, interpretation of the pattern results and applying the knowledge discovered.

2. Methodology

This research applies data mining techniques in a labeled data set i.e. a collection of spam and non-spam emails called SPAMBASE [3], which contains 4, 601 tuples previously identified – 1, 813 classified as non-spam (39.40% of the base) and 2,788 classified as spam (60.60% of the base) e-mails. Weka was used in order to analyze and quantify the types of spam e-mails presented in this data set helping the network management process.

This experiment used a public data set called SPAMBASE, which contains fifty seven data attributes and one classification attribute to determine the type of the content. This data set was created in order to improve security software in computer networks as attacks using spam e-mails can cause losses such as unnecessary time spending, cost increasing, productivity loss, improper or offensive content and financial loss caused by fraud [4].

The UCI Machine Learning Repository has developed the Spambase Database' which is located at its data bed. [5]

2.1 Weka Tool

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License, which contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality [6].

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported) [7] [8].

The Preprocess panel has facilities for importing data from a database, a CSV file, etc., and for preprocessing this data using a so-called filtering algorithm. These filters can be used to transform the data (e.g., turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria [7].

The Classify panel enables the user to apply classification and regression algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, ROC curves, etc., or the model itself.

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes.

2.2 The SPAMBASE Database

To analyze and quantify the anomalies presented in the SPAMBASE data set, data mining techniques were used through Weka. The data set could be loaded into Weka in an Attribute-Relation File Format file (ARFF). In this file, each column contains the type of data, I.e. attribute, for instance: a number or character and in each line there is a data with its respective data delimited by commas.

This research uses the SPAMBASE data set converted in CSV – Comma Separated Values and then converted into ARFF, compatible with Weka data miner. All the fifty-seven attributes – as each attribute represents a word and its frequency in a given email – all the SPAMBASE data set were selected as they have relevant characteristics for the tests.

2.3 The Classifiers

To execute the data mining, Discretize filter was applied first and then two techniques were used separately: (J48) decision tree algorithm and the Artificial Neural Network (ANN) Multi-Layer Perceptron (MLP) commonly used in data mining classification [9] [10].

2.3.1 Decision Tree (J48)

J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. C4.5 is an algorithm used to generate a decision tree, which can be used for statistical classification [11].

C4.5 builds decision trees using the concept of information entropy. The training data is a set of already classified samples. Each sample (Si) consists of a p-dimensional vector, where the Xj represent attributes or features of the sample, as well as the class in which (Si) falls [12].

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recourses on the smaller sub lists [13].

2.3.2 Radial-Basis Function (RBF)

Radial basis function network (RBF) have been successfully employed in different Machine Learning problems. The use of different radial basis functions in RBF has been reported in the literature [24].

RBF network are a class of Artificial Neural Networks where RBF are used to compute the activation of artificial neurons. RBF networks have been successfully employed in real function approximation and pattern recognition problems. In general, RBF networks are associated with architectures with two layers, where the hidden layer employs RBF to compute the activation oh the neurons are a class of real-valued functions where its output depends on the distance between the input and the center of the RBF.

However the author [25] describe that RBF network are class of the Weka tool that implements a normalized Gaussian radial basis function network. It uses the k-means clustering algorithm to provide the basis functions and learns either a logistic regression (discrete class problems) or linear regression (numeric class problems) on top of that. Symmetric multivariate Gaussians are fit to the data from each cluster. If the class is nominal it uses the given number of clusters per class. It standardizes all numeric attributes to zero mean and unit variance [25] [26].

2.3.3 Multi-Layer Perceptron (MLP)

In order to classify, the Artificial Neural Network (ANN) Multi-Layer Perceptron (MLP), was used, with a supervised learning paradigm [14] [15].

The ANN tries to calculate the output layer error and sends the result backwards to the hidden layers to update the weight values in all layers through backpropagation [14].

The training process has two phases: the forward propagation, used to generate the propagation’s output activations of a training pattern’s input and the backward propagation, which uses a given output and the network output to update the weight values in all layers [14]. The MLP network will only be considered trained when the error rate among the outputs is reduced to an accepting value, i.e., this value depends on how the algorithm is used.

According to Haykin [14], the algorithm follow any steps like Initialization, Training sample presentation, Propagation and Backpropagation.

2.4 Standard Statistic Methods

A standard statistic method called Cross Validation [143] was chosen to assess the algorithms. Cross validation divides the data set randomly into a training group formed by 75% of the data set and into a test group formed by 25% of the data set. The samples are then divided into 10 partitions mutually exclusive. The procedure is repeated in each partition. The assessment degree of the classifier will be set this way to guarantee this method.

2.5 Classifier Performance Assessment

A result analyses technique called ROC Curve (*Receiver Operating Characteristic*) [16], developed by Metz [17], has been used in order to assess the classifiers’ performance in similar researches.

This method presents a result that can only show two options, there are abnormalities or not. The tuple has an abnormality or not. It is possible to have an affirmative or negative answer. In an affirmative case, (an abnormality is present) the result is a true positive (TP). It occurs when there is an abnormality and it is possible to notice it; and a true negative (TN), when it is not possible to notice an abnormality [18].

However, it is possible to interpret a normal data as an anomaly and the result will be false positive (FP), or to interpret an anomaly as a normal data and the result will be false negative (FN).

Each point in the ROC Curve represents a different threshold between the fraction of true-positives and the fraction of falsepositives, i.e., the ROC Curve is conceptually similar to a curve, which shows the relation between the test strength and the probability to commit a mistake. Each point in the ROC Curve describes criteria to distinguish a normal data or an anomaly. These are the operation points in the ROC Curve [18].

Precision is an index that indicates the fraction of correctly classified cases. Precision is calculated by sensibility and specificity terms. Sensibility is a parameter, which indicates how many correct positive results occur, and specificity is a parameter, which indicates how many incorrect positive results occur. This precision means enables trust ability of the information [18].

According to Metz [17], a ROC Curve represents the performance, which can be reached between sensibility and specificity in a diagnosing system when the threshold is varied. A comparison between the systems can be done through the areas of the curve in each system [15].

The Az area in the ROC Curve is one of the indexes more frequently used and it represents the correct results in the system (classifier), i.e., the biggest the area, the greater is the correct results. This means that if the system is well weighted and highly precise, the curve should be the nearest possible of the upper left part of the Cartesian ax, increasing the curve area. [19] [20] [21] [22] [23].

3. Results

After completing the data mining using both algorithms, it was possible to achieve a great amount of correctly classified instances using all the SPAMBASE data set with 4, 601 instances. It is important to highlight that the cross-validation method was used to assess the tests.

The J48 algorithm presented a 92.76% rate of correctly classified instances, where 89.79% were classified as non-spam e-mails and 93.34% were classified as spam e-mails according Table 1 and Az equal 0.941.

Spam	Non-spam	Rate%
2602.32	185.68	93.34%
185.11	1627.89	89.79%
	Average rate %	92.76%

Table 1. Confusion matrix – J48

The RBF neural network was set as follow: 57 input attributes, an intermediate, an output layer with 1 neuron, 2 classes. After the RBF data mining was finished, it was possible to obtain 84, 30% rate of correctly classified instances where 89,99% rate of correctly classified non-spam e-mails and 78, 60% rate of correctly classified spam e-mails showing Table 2, Az equal 0.92.

Spam	Non-spam	Rate%
2509	279	89.99%
388	1425	78.60%
	Average rate %	84.30%

Table 2. Confusion matrix – RBF

The MLP neural network was set as follows: 57 input attributes, an intermediate layer with 68 neurons, an output layer with 1 neuron, 0.3 learning rate and 0.2 momentum. After the MLP data mining was finished, it was possible to obtain 93.89/% rate of correctly classified instances where 93.93% rate of correctly classified non-spam e-mails and 93.87% rate of correctly classified spam e-mails showing Table 3, Az 0.98.

Notice that we used the ROC curve (Receiver Operating Characteristics), which is a technique for analyzing the performance of classifiers. The results are promising and we are encouraged to open further studies using the pilot results and outcome.

Spam	Non-spam	Rate%
2617.10	170.90	93.93%
110.05	1702.95	93.87%
	Average rate %	93.89%

Table 3. Confusion matrix – MLP

4. Conclusion

Thus spam detection databases have potential in studying the spam attack. We have used this large dataset SPAMBASE, which presents a collection of spam and non-spam e-mails and applied in this research. In this research, it was proposed to use all the attributes of the data set. After using the Discretize filter and three data mining techniques – J48 decision tree, RBF and MLP Neural Network. It was possible to obtain 92.76% rate of correctly classified instances with the J48 algorithm, 89,99% rate of correctly classified instances with the RBF Neural Network and 93.89% rate of correctly classified instances with the MLP Neural Network. Thus we have presented the benefits of using these techniques to detect spam e-mails in computer networks. It is important to highlight that the MLP Neural Network presented a greater rate of correctly classified instances due to its generalization characteristics, concluding that it is currently the best data mining technique in detecting spam e-mails in this moment.

References

[1] Fayyad, U. M., Piatesky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. *In*: Advances in Knowledge Discovery and Data Mining, AAAI Press.

[2] Narendran, C. R. (2009). Data Mining - Classification Algorithm – Evaluation, May 8th.

[3] Hopkins, M., Reeber, E., Forman, G., Suermondt, J. SPAMBASE, http://www.ics.uci.edu/~mlearn/databases/spambase/

[4] Fogel, J., Shlivko, S. (2010). Weight Problems and Spam E-mail for Weight Loss Products Southern Medical Journal, v. 103, ed. 1, p 31-36.

[5] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/spambase/

[6] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (2009). The Weka Data Mining Software: An Update; SIGKDD Explorations, 11 (1).

[7] Witten, I. H., Frank, E., Hall, M. A. (2011). Data Mining: Practical machine learning tools and techniques, 3rd Edition. Morgan Kaufmann, San Francisco. Retrieved.

[8] Holmes, G., Donkin, A., Witten, I. H. (1994). Weka: A machine learning workbench. Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia. Retrieved 2007-06-25.

[9] Ribeiro, P. B., Schiabel, H., Romero, R. A. F. (2010). Comparativo entre Classificadores de Nódulos Mamários. In: XXII Congresso Brasileiro de Engenharia Biomédica (CBEB, 2010), Tiradentes, MG. Anais do XXII CBEB - 2010 (ISSN: 2179-3220). p.165 – 168.

[10] Ribeiro, P. B., Schiabel, H., Romero, R. A. F. (2011). Artificial Neural Networks versus Systems Fuzzy in Breast Masses Classification Schemes. *In*: Society for Imaging Informatics in Medicine (SIIM) - Junho 2-5/2011, Washington, DC.Society for Imaging Informatics in Medicine (SIIM).

[11] Karimi, K., Hamilton, H. J. (2002). TimeSleuth: A Tool for Discovering Causal and Temporal Rules, ICTAI.

[12] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

[13] Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques, *Informatica*, 31, 249-268,

[14] Haykin, S. (2009). Neural Networks and Learning Machines. Editora Prentice Hall, 3a. Edição, p. 936.

[15] Silva, I. N., Spatti, D. H., Flauzino, R. A. (2010). Redes Neurais Artificiais: para engenharia e ciências aplicada. Ed. Artliber, p. 399.

[16] Suri, J. S., Rangayyan, R. M. (2006). Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer. Bellingham, Washington, SPIE Press.

[17] Metz, C. E. (1986). ROC Methodology in Radiologic Imaging. Investigative Radiology, 21, p. 720–733.

[18] Evans, A. L. (1981). The evaluation of medical images. Adam Hilger Ltd, Bristol, Great Britain.

[19] Dorfman, D., Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. *Journal of Mathematical Psychology*, 6, 487–496.

[20] Dorfman, D., Rscore, J., Pickett, R. M. (1982). Evaluation of diagnostic systems: Methods from signal detection theory. New York: Academic Press, 212-232.

[21] Dorfman, D., Berbaum, K.S., Rscore, J. (1986). Pooled rating-method data: a computer program for analyzing pooled ROC curves. Behavior Research Methods, *Instruments, and Computers*, 18, 452-462.

[22] Hanley, J. A., Mcneil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143, p. 29-36.

[23] Hanley, J. A., Mcneil, B. J. (1983). A method of comparing the areas under receiving operating characteristic curves derived from the same cases. Radiology, 148, p. 839-843.

[24] Silva, L. E. V., Duque, J. J., Tinós, R., Murta Jr, L. O. (2010). Reconstruciton of Multivariate Signals Using Q-Gaussian Radial Basis Function Network. Computing in Cardiology, 37, p. 465-468.

[25] Khade, G., Kumar, S., Bhattacharya, S. (2012). Classification of Web on Attractiveness: A Supervised Learning Approach. IEEE Procedings of 4th International Conference on Intelligent Human Computer Interaction, p. 27-29.

[26] Moodly, J., Darken, C. J. (1989) Fast Learning in Networks of Locally Tuned Processing Units. Neural Computation, 1, p. 281-294.

Author Biographies

Kelton Costa received his B.Sc. in Systems Analysis from Sagrado Coração University. He received: in 2004 M.Sc. Computer Science from the University Center Eurípides. In 2009 Ph.D. Electrical Engineering from the São Paulo State University. During 2010-2013, he worked as a post-doctorate researcher at the Institute of Computing of the University of Campinas and in the Department of Computer Science of the Paulista Júlio de Mesquita Filho State University. He is a professor in the Department of Computing, College of Technology of the São Paulo State and his research interests include security in computer networks and detecting anomalies.

Patricia Bellin Ribeiro received her Ph.D. degree in science (2013) and Master’s degree in Electrical Engineering (2006), area of concentration Signals and Image Processing, by School of Engineering of São Carlos of University of São Paulo. In Ph.D. her worked with segmentation for automated analysis of mass in digital mammograms, applied for CADx and Master’s degree he worked with the classification of medical images. She is a professor in the Department of Computing, College of Technology of the São Paulo and her research interests include artificial neural networks, fuzzy logic, image processing.

Atair Alves Camargo Junior is currently studying Computer Networks at FATEC (Faculdade de Tecnologia - Bauru in Technological College). He has been researching Data Mining and Neural Networks since 2012. He also researched pattern recognition during his first college degree. He has worked as a teacher of English Language for 13 years. He also worked as a network analyst at Brasil Telecom / Oi Telecom in São Paulo in Brazil.

Victor Vavali Rossi is studying Database at FATEC (College of Technology). He has been researching Data Mining and Neural Networks since 2012. He also researched pattern recognition during his first college degree.

Henrique Martins is graduated in Analysis Systems at Sagrado Coração University in Bauru, São Paulo, Brazil, in 2001, specializes in Information Systems and Master student Computer Science in the Department of Computer Science of the Paulista Júlio de Mesquita Filho State University. His current research interest in Computer Networks, Distributed Systems and Cloud Computing.

Miguel Neves is Master in Design and Technology Intelligent Digital of Catolic the Pontifical University of São Paulo, Degree in Systems Information Technologies at College of Technology in Ourinhos. He is currently professor at Fatec Bauru. Has experince in Computer Science in programming languages, Database analysis of Systems, Software Engineering, database and hypermedia.

Ricardo Fabris is student in Computer Networks at the College of Technology of the São Paulo State, Bauru, São Paulo, Brazil. His current research interest in Computer Networks, Distributed Systems, Cloud Computing and Artificial Intelligence.

João Paulo Papa received his B.Sc. in Information Systems from the São Paulo State University, SP, Brazil. In 2005, he received his M.Sc. in Computer Science from the Federal University of São Carlos, SP, Brazil. In 2008, he received his Ph.D. in Computer Science from the University of Campinas, SP, Brazil. During 2008-2009, he had worked as post-doctorate researcher at the same institute. He has been Professor at the Computer Science Department, São Paulo State University, since 2009, and his research interests include machine learning, pattern recognition and image processing.