FEATURE SELECTION, LEARNING METRICS AND DIMENSION REDUCTION IN TRAINING AND CLASSIFICATION PROCESSES IN INTRUSION DETECTION SYSTEMS

[1]FABIO MENDOZA PALECHOR, [2]ALEXIS DE LA HOZ MANOTAS, [3]EMIRO DE LA HOZ FRANCO, [4]PAOLA ARIZA COLPAS
[1234] Associate Professor, Systems Engineering Department, Universidad de la Costa, Colombia
E-mail: [1]fmendoza1@cuc.edu.co, [2]adelahoz6@cuc.edu.co, [3]edelahoz@cuc.edu.co, [4]pariza1@cuc.edu.co

ABSTRACT

This research presents an IDS prototype in Matlab that assess network traffic connections contained in the NSL-KDD dataset, comparing feature selection techniques available in FEAST toolbox, refining prior results applying dimension reduction technique ISOMAP. The classification process used a supervised learning technique called Support Vector Machines (SVM). The comparative analysis related to detection rates by attack category are conclusive that MRMR+PCA+SVM (selection, reduction and classification techniques) combined obtained more promising results, just using 5 of 41 available features in the dataset. The results obtained were: 85.42% normal traffic, 80.77% DoS, 90.41% Probe, 91.78% U2R and 83.25% R2L.

Keywords: *System Intrusion Detection (IDS), Feature Selection Toolbox (FEAST), Isometric Feature Mapping ISOMAP, Support Vector Machine (SVM), Principal Component Analysis (PCA).*

1. INTRODUCTION

The diverse Internet topology networks are exposed to a growing number of security threats. With new kind of attacks appearing permanently, the development of adaptive security approaches to the recurring attack changes, has become one of the main focus in research **[1]**. The need of constant adaptation to IDS (Intrusion Detection Systems) to a growing and wider attack set, has led to integration of artificial intelligence based techniques to IDS, becoming a strong tendency of research interest for scientific community **[2]**.
The 2015 Internet Security Threats report from Symantec **[3]**, show relevant increase in vulnerabilities identification, from 127 in 2013 to 168 in 2014, mail phishing campaigns from 779 to 841 in the same period and security flaws from 253 to 312, which represents a 23% increase and 348 million exposed to possible fraudulent actions.

Cisco Annual Security Report **[4]** indicates that client side vulnerabilities are the main focus for attackers such as Adobe Flash Player, Internet Explorer and Apache Struts framework, with a correspondent increase in the attacks of malware and SSL certification fraud.

An alternative to intrusion detection is the use of data mining tools and techniques that facilitate potential risks identification in computer network traffic. This research applied feature selection techniques in **FEAST** Matlab toolbox, combined with ISOMAP reduction dimension technique and Support Vector Machines (SVM) as supervised learning method. After this, an IDS prototype in Matlab was obtained, that evaluates network traffic connections contained in the NSL-KDD dataset.

Based on the final results, using the data mining techniques previously mentioned, it can be identified intrusion detection techniques with higher precision rates and effectiveness, which leads to bigger benefits for institution and personal network users.

2. RELATED WORK

The exponential behavior that has been normal in Internet has highlighted security flaws in the network protocols implementations. This situation is understandable since most network protocols were originally designed to communication scenarios with few machines (hosts), which it's ratified with the operation of the All-IP networks, which work by the precept of the best-effort. **[5]** Facts like these, has turned network security in attractive research area.

Through time, there have been many technologies capable of identifying effectively, perpetrated attacks to computer networks, which has allowed the creation of different mitigation strategies, being one of the most efficient, intrusion detection systems (IDS), which nowadays has led to a new research area called " Intrusion Detection based on anomalies", proposing the use of different techniques such as: artificial intelligence, data mining and machine learning. Next, a brief review of related studies with the objective of finding tendencies in this research area is presented.

- In **[6]** a methodology for IDS efficiency validation is their objective, and propose three phases (selection, training and classification), using FDR as feature selection technique, self organizing maps (SOM) for classification. The obtained results show 97,39% sensitivity and 62,73% specifity using 17 features of the dataset.
- In **[7]** an implementation of the algorithms found in the FEAST Matlab toolbox with the goal of selecting the best method based on accuracy using the least number of features. The dataset used was NSL-KDD. The RELIEF method obtained the best accuracy results on attack detection: 86,20% (NORMAL), 85,71%(DoS), 88,42%(PROBE), 93,11% (U2R) and 90,07(R2L), which makes it a promising technique for feature selection applied to intrusion detection processes on computer networks.
- In **[8]** Shows a dimension reduction method PCA for the data preparation and then implements a neural classifier achieving 99.3% as result in attack identification. The entry data is a vector of 419 variables reduced to only 20 variables.
- In **[5]** genetic algorithms are implemented for normal traffic pattern recognition obtaining promising results given coincidences for the dataset in 80%.
- In **[9]** shows the design of an intruder detection method based on SVM (Support Vector Machines) looking for attacks identification in computer networks obtaining 94% of precision.
- In **[10]** an instruction detection method is built, using the CSI-KNN K-closing neighbors algorithm. The algorithm analyzes the different features of data of the network using two measures: strangeness and isolation. Based on these measures, a correlation unit increase intrusion alerts with the trust estimation associated.
- In **[11]** data from DARPA 1998 are used to study the intruder detection systems, comparing performance of Robust Support Vector Machines (RVSMs) against Support Vector Conventional machines and classifier of the nearest neighbor. The results indicate the RSVMs superiority in terms of high accuracy in intrusion detection and few false positives, also in their capacity for generalization in the presence of noise and working time.
- In **[12]** a new framework based in data mining techniques for IDS design is proposed, and the classification process is based in association. The classification algorithm process proposed uses fuzzy association rules for construction classifiers. A new method to accelerate the induction rules algorithm through element reduction is also proposed. The used dataset is KDD-99. The results of the attacks weren't that promising, nevertheless the total detection rate of known attacks was quite high, oppose to the the false positive one. The detection rate achieved was 80,6% and the false positive rated was 2,95%.
- In **[13]**, a new approach is proposed called FC-ANN, based on ANN and Fuzzy Clustering to solve the problem and help the IDS to obtain a greater detection rate, lesser false positive rate and stronger stability. The dataset used was KDD CUP 1999. The obtained results in precision were FC-ANN(96,71%), BPNN (96,65%), Decision Trees(96,75%), NaBayes(96,75%).

3. FEATURE SELECTION AND PROCESSING

Feature selection is a data mining technique with the objective of reducing the entry data dimension. It´s definitely a significant process, which involves rejecting those attributes that aren't relevant enough to go through data analysis **[14]**.

A dataset can contain "n" columns that describe features of computer network connections, but if some of their data is too disperse, it wouldn't be useful or any advantage add them to the model. If any unnecessary columns are added through the model generation, it would take more processing and memory during training stage, and storage space to save the complete model.

A typical feature selection process can be executed in 4 steps **[15]**:

- **Subset Generation**, it's a searching process to obtain a subset of available features to be evaluated. It can be exhaustive **[16]** or sequential **[17]**.
- **Subset Evaluation**, verifies optimal level of the subset previously created to be implemented in a learning problem, which in this case is a classification problem.
- **Stopping Criteria**, decides when a feature selection process should be stopped.
- **Result validation**, it evaluates the feature selection model with real data to verify its performance against experimental results.

One of the problems found in feature selection processes makes reference to the election of a single specific method, due to the fact that it's unknown which one could offer the best classification for intrusion detection **[2]**.

Due to the complexity of processing big volumes of data, several methods and algorithms have been proposed that allow optimal feature selection for data processing. In this research, the feature selection methods found in FEAST Matlab toolbox are used **[18]**.

FEAST provides implementation of feature selection algorithms based on common information. All functions require discrete inputs and return selection function indexes. It includes implementation of mim, mrmr, mifs, cmim, jmi, disr, cife, icap, condred, cmi, relief, fcbf and betagamma methods **[18].**

Dimension reduction is a very important aspect when big volumes of data are processed since the number of variables is huge so it can be improved using dimension reduction methods found in Matlab toolbox, which contains 34 dimension reduction techniques and learning metrics, some of them are Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), ISOMAP and Kernel PCA **[19]**.

3.1 Principal Component Analysis (Pca)

It´s a data pattern recognition technique **[20]**, looking for highlight similarities and differences between them. Due to the large amount of data processed, pattern recognition it's quite difficult whereas graphical representation it`s not possible, so PCA becomes a powerful tool for data analysis.
Another advantage of PCA is that after finding the patterns, these are compressed, using dimension number reduction, without any loss information involved. This technique is widely used in image compression.
According to **[21]**, PCA has been used in many applications for relevant data extraction. In fact, it has become a total success and has been implemented in facial recognition applications **[22].**

3.2 Kernel Principal Component Analysis (KPCA)

KPCA **[23]** is a non linear technique for feature extraction, closely related to Vector Support Machines. It has demonstrated to be quite useful for several applications such as noise reduction **[24]** and pre-processing in regression problems **[25]**.

3.3 Isometric Feature Mapping (ISOMAP)

It's a non linear method for dimensionality reduction. Its purpose is to search for the map holding the total non linear geometry of the data trough geodesic preservation (shorter curve above the collector connecting two points). First, makes a calculation of the distances between geodesic points, then execute MDS (Multidimensional Scaling) to find the projection that keeps those distances. According to **[26]** ISOMAP it´s based in comprehensive analysis of the data features with the purpose of reducing dimensionality.

3.4 Support Vector Machines (SVM)

It´s a last generation classification technique that has become relevant in recent years **[27] [28]**. SVM is based on the structural risk reduction **[29]**. In many applications, SVM has demonstrated to obtain proper results, surpassing learning machines such as neural networks **[28]** and has been used as efficient tools for solving classification problems.

Researchers as **[30]**, mention that Support Vector Machine learns the decision surface of two different classes of entry points. Since it´s a one class classifier, the description given by the data of the support vectors is capable of building a decision frontier around the learning data domain with few or no knowledge whatsoever of the data outside this frontier. The data are mapped with a Gaussian kernel or other kind of kernel to a feature space in a higher dimensional space, where maximum separation between classes is looked for. This frontier function, when brought back to the entry space, can separate the data in different classes, each of them forming a group.

4. DATASETS

NSL-KDD it´s a Dataset that solves some problems inherent to KDD'99 **[31]**. This new version of the KDD data, still suffers from some problems and it's not a perfect representation of real existing networks due to lack of public datasets for IDS. Nevertheless, it can be used as a efficient reference of real network traffic, to help research that makes comparisons between intrusion detection methods. On the other side, the record number of NSL-KDD are quite reasonable. This advantage makes it accessible to perform experimentation using the whole dataset without having to select randomly a subset of it. In consequence, the evaluation results of the research based on it will be consistent and comparable.

Since 1999, KDD'99 has been the most widely used dataset for evaluation of anomaly detection methods. This dataset was collected by **[32]**, and it´s built over the database captured in the evaluation program IDS DARPA'98. DARPA'98 is a dataset resulting of 7 weeks of network traffic, that can be processed in almost 5 millions of connection records, each one of them with near 100 bytes.

The test data represent two weeks and have almost 2 million connection records. KDD data is composed approximately of 4.900.000 connection vectors each of them with 41 features and tagged as normal or a specific attack. The attacks can ben Denial of Service "DoS", "Probe", Remote to Local "R2L" and User to Root "U2R" **[33].**

Next, the 41 features present in each connection record of the dataset NSL-KDD.

No.	FEATURE NAME
1	duration
2	protocol
3	service
4	flag
5	Src bytes
6	Dst bytes
7	Land
8	wrong fragment
9	Urgent
10	Hot
11	num failed logins
12	logged in
13	num compromised
14	root Shell
15	su attempted
16	num root
17	num file creations
18	num shells
19	num access files
20	num outbound cmds
21	is host login
22	is guest login
23	Count
24	srv coun
25	serror rate
26	srv serror rate
27	rerror rate
28	srv rerror rate
29	same srv rate
30	diff srv rate
31	srv diff host rate
32	dst host count
33	dst host srv count
34	dst host same srv rate
35	dst host diff srv rate
36	dst host same src port rate
37	dst host srv diff host rate
38	dst host serror rate
39	dst host srv serror rate
40	dst host rerror rate
41	num outbound cmds
42	Class

5. METHODOLOGY

This research is based on the use of the NSL-KDD dataset, FEAST as feature selection method and ISOMAP, PCA and KPCA as dimension reduction methods and learning metrics. Next, the stages of the present study.

- First, a normalization process was applied to the dataset NSL-KDD after being downloaded from its source.
- Next, the Matlab R2013a settings were tuned to be able to execute the feature selection,

learning metrics and dimensionality reduction methods.

- The dataset distribution was made recreating two different simulation scenarios, based on cross validation: first scenario 20% of the data was used for training and 100% of data was used on testing; in the second scenario, 100% of data was used for training and testing.
- Next, the experimental phase was executed, using the feature selection methods JMI, MRM and RELIEF found in FEAST, with ISOMAP, PCA and Kernel-PCA techniques, each experiment was made taking 5 features of the dataset. This allowed to present a purpose of attack detection taking into account the best feature selection techniques, learning metrics and dimensionality reduction methods.

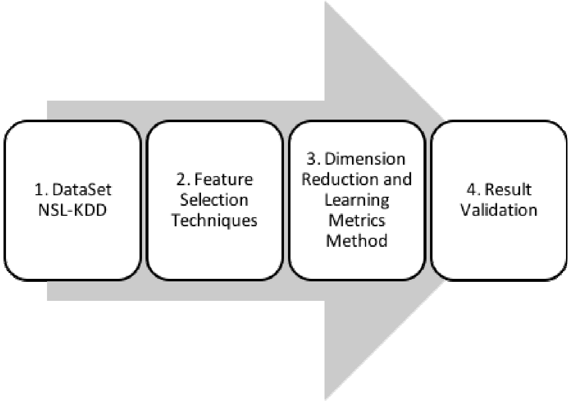
1. DataSet NSL-KDD
2. Feature Selection Techniques
3. Dimension Reduction and Learning Metrics Method
4. Result Validation

Figure 1: Intrusion Detection Prototype

- Finally, after experimentation concluded, results were compared based on the precision performance of each method.

6. Results

The classification process was developed using Matlab R2013a. The data used were:

a. KDDTrain + 20 %; 41 attributes; 25.192 records.
b. KDDTrain + 100 %; 41 attributes; 125.973 records
c. KDDTest + 100 %; 41 attributes; 22.544 records

With the defined dataset, the best three feature selection methods were used from FEAST [7] with dimension reduction and learning metrics techniques PCA, ISOMAP and Kernel-PCA. As a result of this comparison, tables and graphs with the precision data of each method were built, which makes it easier to determine the best techniques with stronger detection level in the different attacks contained in NSL-KDD dataset.

Table 2: Precision Values Of FEAST Methods With Dimension Reduction Techniques And Learning Metrics For Normal Connections

NORMAL ATTACKS			
	PCA	ISOMAP	KPCA
JMI	78,49%	39,32%	53,66%
MRMR	85,42%	62,74%	56,92%
RELIEF	76,40%	65,02%	43,12%

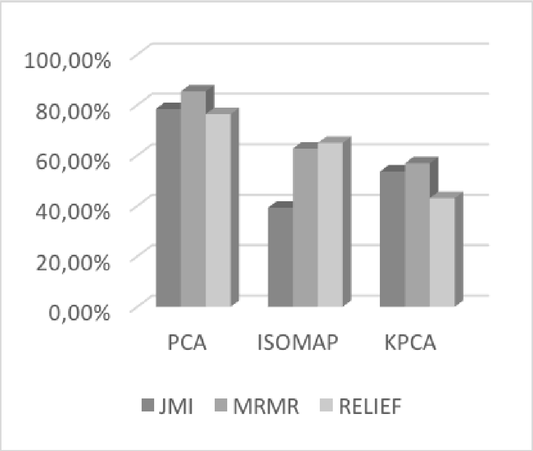

Figure 2: Precision Values Of FEAST Methods With Dimension Reduction Techniques And Learning Metrics For Normal Connections.

Table 3: Precision Values Of FEAST Methods With Dimension Reduction Techniques And Learning Metrics For Dos Connections.

DOS ATTACKS			
	PCA	ISOMAP	KPCA
JMI	80,19%	82,47%	81,82%
MRMR	80,77%	70,06%	82,27%
RELIEF	85,87%	39,75%	33,37%

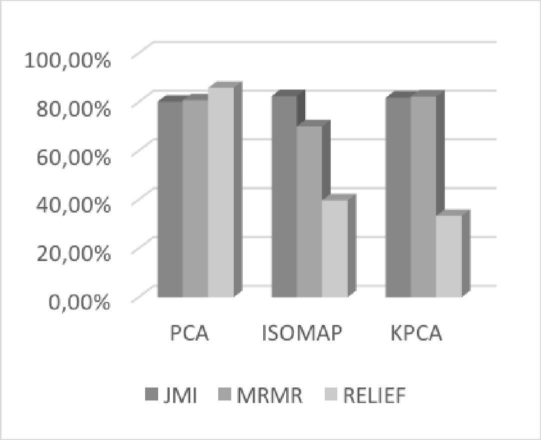

Figure 3: Precision Values Of FEAST Methods With Dimension Reduction Techniques And Learning Metrics For Dos Connections.

Table 4: Precision Values Of FEAST Methods With Dimension Reduction Techniques And Learning Metrics For PROBE Connections.

PROBE ATTACKS			
	PCA	ISOMAP	KPCA
JMI	92,56%	53,21%	90,57%
MRMR	90,41%	91,75%	93,31%
RELIEF	82,60%	69,07%	13,19%

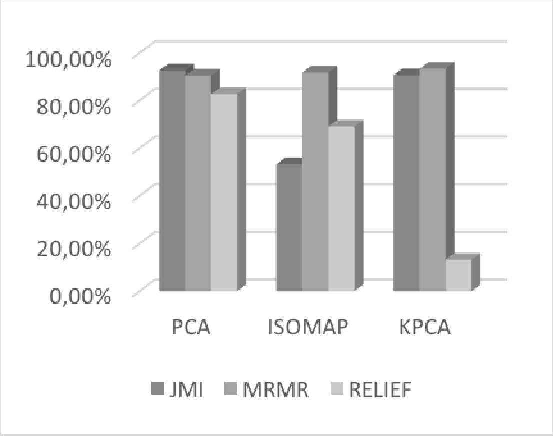

Figure 4: Precision Values Of FEAST Methods With Dimension Reduction Techniques And Learning Metrics For PROBE Connections.

Table 5: Precision Values Of FEAST Methods With Dimension Reduction Techniques And Learning Metrics For U2Rl Connections.

U2R ATTACKS			
	PCA	ISOMAP	KPCA
JMI	91,87%	57,07%	96,55%
MRMR	91,87%	1,41%	99,60%
RELIEF	56,40%	66,35%	99,70%

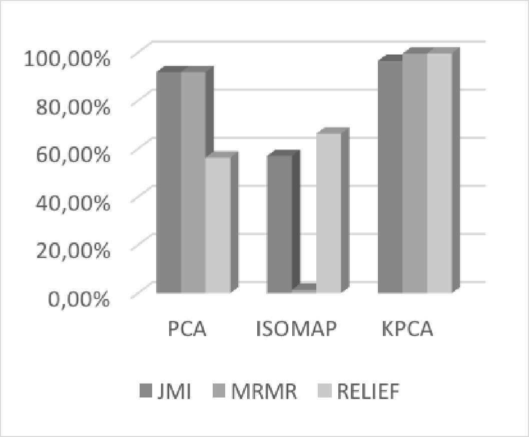

Figure 5: Precision Values Of FEAST Methods With Dimension Reduction Techniques And Learning Metrics For U2R Connections.

Table 6: Precision Values Of FEAST Methods With Dimension Reduction Techniques And Learning Metrics For R2L Connections.

R2L ATTACKS			
	PCA	ISOMAP	KPCA
JMI	83,25%	86,87%	87,19%
MRMR	83,25%	86,87%	87,19%
RELIEF	78,40%	60,63%	87,13%

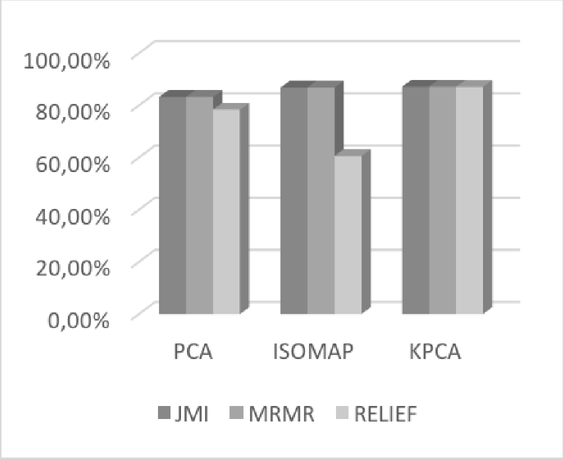

Figure 6: Precision Values Of FEAST Methods With Dimension Reduction Techniques And Learning Metrics For R2L Connections.

7. CONCLUSIONS

The intrusion detection systems constitute great benefit to guarantee the security of different type of networks. Using feature selection, dimension reduction and learning metrics techniques has become a big contribution to the construction of IDS capable of identifying different kind of attacks efficiently.

It's important to know that dimension reduction and learning metrics techniques used to build the IDS were subject of test like ISOMAP, PCA an KPCA, implemented with the best feature selection method MRMR to achieve a system with the best accuracy results for identification of computer network attacks.

The MRMR feature selection method and the PCA dimension reduction and learning metrics method had promising results for identification of different kind of attacks, using only 5 of 41 possible features in the dataset. The results were: 85,42% (NORMAL), 80.77%(DoS), 90.41% (Probe), 91,87% (U2R) and 83,25% (R2L).

REFRENCES:

[1] Garcia, P., Diaz, J., Macia, G. and Vasquez, E., "Anomaly-based network intrusion detection: Techniques, systems and challenges", in journal Computers & Security, Vol. 28, pp. 18-28, 2009.

[2] Xiaonan, S. and Banzhaf, W., "The use of computational intelligence in intrusion detection systems: A review", in journal Applied Soft Computing, Vol. 10, pp. 1-35, 2010.

[3] Symantec. 2015 Internet Security Threat Report [online]. Available: http://www.symantec.com/security_response/publications/threatreport.jsp

[4] Cisco Systems. Cisco survey evolving security threats [online]. Available: http://www.enterprisetech.com/2015/04/07/cisco-survey-sees-evolving-security-threats/

[5] Catania, C., Garcia, C., "Reconocimiento de Patrones en el Trafico de Red Basado en Algoritmos Genéticos", Revista Iberoamericana de Inteligencia Artificial, Vol 12, pp. 65-75, 2008.

[6] De la hoz, E., Ortiz, A., Ortega, J., De la hoz, E. And Mendoza, F., "Implementation of an Intrusion Detection System Based on Self Organizing Map", in Journal of Theoretical and Applied Information Technology, Vol. 71, pp. 324-334, 2015.

[7] Mendoza, F., De la hoz, E. And De la hoz, A., "Application of Feast (Feature Selection Toolbox) in IDS (Intrusion Detection Systems)", in Journal of Theoretical and Applied Information Technology, Vol. 70, pp. 579-585, 2014.

[8] Lorenzo, I., Macia, F., Mora, F., Gil, J., and Marcos, J., "Modelo Eficiente y Escalable para la Deteccion de Intrusos en Red", in XXIV Simposium Nacional de la Unión Científica Internacional de Radio (URSI'09), 2009.

[9] Xiaoqing, G., Hebin, G., and Luyi, C., "Network Intrusion Detection Method Based on Agent and SVM", in Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on, pp. 399 – 402, 2010.

[10] Kuang, L., and Zulkernine, M., "An Anomaly Intrusion Detection Method Using the CSIKNN Algorithm", in Proceedings of the 2008 ACM Symposium on Applied Computing, pp. 921-926, 2008.

[11] W. Hu, Y. Liao, and V. Vemuri. "Robust Support Vector Machines for Anomaly Detection in Computer Security". In ICMLA, pp. 168–174, 2003.

[12] Tajbakhsh, A., Rahmati, M., and Mirzaei, A., "Intrusion detection using fuzzy association rules". In Applied Soft Computing, Vol. 9(2), pp. 462-469, 2009.

[13] Wang, G., Hao, J., Ma, J., and Huang, L., "A new approach to intrusion detection using

Artificial Neural Networks and fuzzy clustering". In Expert Systems with Applications, Vol. 37(9), pp. 6225-6232, 2010.

[14] Microsoft. Selección de Características (Minería de Datos) [online]. Available: https://msdn.microsoft.com/es-es/library/ms175382(v=sql.120).aspx

[15] Oporto, S., Aquino, I., Chavez, J., Perez, C., Comparación de Cuatro Técnicas de Selección de Características Envolventes usando Neuronales, Arboles de Decisión, Maquinas de Vector de Soporte y Clasificador Bayesiano.

[16] Goldberg, D. And Holland, J., "Genetic algorithms and machine learning", in Machine learning, Vol. 3(2), pp. 95-99, 1998.

[17] Liu, H., & Motoda, H, "Feature selection for knowledge discovery and data mining", in Springer Science & Business Media, Vol. 454, 2012.

[18] University of Manchester. A Feature Selection Toolbox for C and Matlab [online]. Available: http://www.cs.man.ac.uk/~gbrown/fstoolbox/.

[19] Van Der Maaten, L., Matlab Toolbox for Dimensionality Reduction [online]. Available : http://lvdmaaten.github.io/drtoolbox/

[20] Lohweg, V., and Mönks, U., "Fuzzy-Pattern-Classifier Based Sensor Fusion for Machine Conditioning". INTECH Open Access Publisher, 2010.

[21] De la Hoz, E., De La Hoz, E., Ortiz, A., Ortega, J., and Prieto, B., "PCA filtering and probabilistic SOM for network intrusion detection", in Neurocomputing, vol. 164, pp. 71-81 2015.

[22] Turk, M. and Pentland, A., "Eigenfaces for Recognition", in journal of cognitive neuroscience, Vol. 3, pp. 71-86, 2007.

[23] Schölkopf, B., Smola, A., and Müller, K., "Nonlinear component analysis as a kernel eigenvalue problema", in Neural computation, Vol. 10(5), pp. 1299-1319, 1998.

[24] Mika, S., Schölkopf, B., Smola, A. J., Müller, K. R., Scholz, M., and Rätsch, G., "Kernel PCA and De-Noising in Feature Spaces". In NIPS, Vol. 4, No. 5, pp. 7, 1998.

[25] Rosipal, R., Girolami, M., and Trejo, L., "Kernel PCA for feature extraction and denoising in nonlinear regression". Technical Report No. 4, Department of Computing and Information Systems, University of Paisley, 2000.

[26] Xiao, X., and Tao, C., "ISOMAP Algorithm-Based Feature Extraction for Electromechanical Equipment Fault Prediction", in Image and Signal Processing, 2009. CISP '09. 2nd International Congress on, pp. 1-4, 2009.

[27] Burges, C., Schölkopf, B. And Smola, A., "Advances in kernel methods: Support vector machines". Cambridge, MA: MIT Press, 1999.

[28] Burges, C., "A tutorial on support vector machines for pattern recognition". Data Mining and Knowledge Discovery, vol. 2, no. 2, 1998.

[29] Vapnik, V., "The nature of statistical learning theory". New York: Springer-Verlag, 1995.

[30] Betancourt, G., Las Maquinas de Soporte Vectorial (SVMs), Universidad Tecnológica de Pereira, 2005.

[31] University of California. The UCI KDD Archive [online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[32] MIT Lincoln Laboratory. 1998 DARPA Intrusion Detection Evaluation Data Set. [online]. Available: http://www.ll.mit.edu/ideval/data/1998data.html

[33] Stolfo, S., Fan, W., Lee, W., Prodromidis, A., and Chan, P., "Costbased modeling for fraud and intrusion detection: Results from the jam project," discex, vol. 02, pp. 1130, 2000.

[34] Tribak, H., "Análisis Estadístico de Distintas Técnicas de Inteligencia Artificial en Detección de Intrusos". Tesis Doctoral, 2012.

[35] Sabnani, S., Computer Security: A machine learning Approach, Technical Report, University of London, 2008.