AN INTRUSION DETECTION ALGORITHM FOR AMI SYSTEMS BASED ON SVM AND PCA

Sara Pourfallah[1], Amir H. Jafari[2], Hadi S. Shahhoseini[3], Mitra oleyaeyan[4]

[1,4]Elearning Center, Iran University of science and Technology, Tehran, Iran
[2,3]Electrical Engineering Department, Iran University of science and Technology, Tehran, Iran

ABSTRACT

Nowadays, using the smart metering devices for energy users to manage a wide variety of subscribers, reading devices for measuring, billing, disconnection and connection of subscribers' connection management is an important issue. The performance of these intelligent systems is based on information transfer in the context of information technology, so reported data from network should be managed to avoid the malicious activities that including the issues that could affect the quality of service the system. In this paper for control of the reported data and to ensure the veracity of the obtained information, using intrusion detection system is proposed based on the support vector machine and principle component analysis (PCA) to recognize and identify the intrusions and attacks in the smart grid. Here, the operation of intrusion detection systems for different kernel of SVM when using support vector machine (SVM) and PCA simultaneously is studied. To evaluate the algorithm, based on data KDD99, numerical simulation is done on five different kernels for an intrusion detection system using support vector machine with PCA simultaneously. Also comparison analysis is investigated for presented intrusion detection algorithm in terms of time - response, rate of increase network efficiency and increase system error and differences in the use or lack of use PCA. The results indicate that correct detection rate and the rate of attack error detection have best value when PCA is used, and when the core of algorithm is radial type, in SVM algorithm reduces the time for data analysis and enhances performance of intrusion detection.

KEYWORDS

Intelligent System AMI, intrusion detection systems, support vector machines, PCA

1. INTRODUCTION

Today, the management of energy networks, including control activities, customer invoice and management at peak hours, the use of smart grid power distribution network is of utmost importance. For this purpose Advanced Metering Infrastructure (AMI) and integrated systems including hardware, software, network and designed communication platform by considering information such as consumption, demand, voltage, current will help to better manage the network. This system creates two-way communication platform capable to reading, tuning, monitoring and remote control of the meters, collect, manage, process and analyse the collected data and produce graphs and reports required. Automatically perform all the processes [1-2].

This project done in France and Italy, according to research and engineering consulting institute Zpryme, the number of smart meters installed in the United States of America from 2.47 million in 2007 to 37.29 million in 2011, has grown that large part by corporations leading such as Pacific gas and Electric (PGE), Florida power and light (FPL) and southern California Edison company

DOI: 10.5121/ijci.2014.3401

installed. Installing smart meters with a 97% annual growth is predicted this amount will reach at the end of 2013 to 61.77 million meters [3].

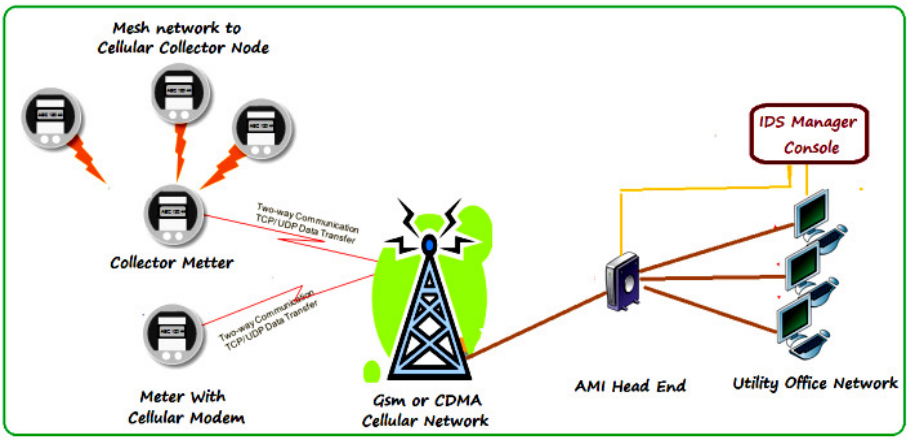

Figure 1: Structure of IDS in the AMI

As can be seen in Figure 1 smart networks using telecommunications equipment and the context of information technology Star, that this area efficiently monitoring and control to complete protective measures and maintain security infrastructure is a critical need. Security in smart system throughout the measurement process from meter and DC to CAS, which are consists many of hardware and software systems must considered and all factors such as manufacturers, suppliers and regulators to increase awareness and ensure security measurement systems will participate together in the future. The following elements can be considered for AMI [4-6]:

- Sensor: hardware or software components or systems for the analysis of network activity. In the case of AMI, sensors should be located at the head-end termination. The sensor head-end termination processes large volumes of traffic; sensors in the meters shall have minimum computing requirements.
- Server management: management of data generated by sensors needs to be sent to one or several servers.
- Database server: store for events information recorded by sensors and server management. A combination of management server and database server that is often Security Information and Event Management (SIEM) is called.
- Console: Interface that security managers can use to 1) configure intrusion detection systems, 2) to monitor the security situation in AMI 3) to visualize and explore the alert, and 4) to perform forensic activities

One of the things that can contributed to the security of these systems is the use of intrusion detection systems in AMI In order to control the traffic these networks be prevented of potential attacks that can be achieved through mesh networks in addition backhaul IP-based networks, imposing to system. This system can be used to identify and deal with these types of attacks that may happen in AMI network. In works [7-9], support vector machine and PCA is proposed but base on our knowledge effect of different SVM kernels in performance of an intrusion detection algorithm when using SVM and PCA simultaneously is not studied. So, in this paper, five kernels of SVM in intrusion detection algorithm that can be used in intelligent network structure such as AMI, by taking advantage of the PCA is explored and assessed by exploiting standard data KDD99 attacks.

In the next section introduces the IDS and its use in identification of attacks would be considered, in the third part, after pre-processing methods, and how to use the support vector machine is presented. The fourth section the proposed is evaluated for standard attack data and the conclusions are presented in Section Five.

2. INTRUSION DETECTION SYSTEMS

Intrusion Detection System (IDS) is responsible for identifying and detecting any unauthorized use of the system, Abuse or damage by both internal and external users [10]. Intrusion detection systems have been created as software and hardware systems and each has its own advantages and disadvantages. Speed and accuracy are the benefits of hardware systems and the lack of security breach by hackers is another the capability of such systems. But the ease of use of the software, the ability to adapt the software requirements and between different operating systems, software systems will be more common and generally these systems have better selection [11].

Generally, three main functions(IDS)are: 1)Monitoring and Evaluation 2) Discovered 3) Reactions ,Thus each IDS can be classified based on intrusion detection techniques, architecture and the response to intrusion and several methods have been designed as intrusion detection techniques to act monitor events occurring in a computer system or network assume.

3. PRE-PROCESSING

To make the data comparable and to be without unit is applied of linear transformation. Also, since the number of attack data features was large and the processing time takes much time is used of PCA to reduce dimension. Principal component analysis (PCA), is a method of reduction dimension, that is based on the work of Pearson. The main goal is, feature extraction has been representing the data in a lower dimensional space with relatively less attention of feature selection. Geometry can be said to PCA, the new vertical axes of the original coordinate axes to be sorted out if the initial variance. Facts do PCA, in Fig2 Is shown. Because of the limitations of the paper is avoid describes the algorithm and reference [12] is presented.

4. SUPPORT VECTOR MACHINES ALGORITHM

During the designing with training data imposed the test set to the model and with calculated error of model in training and testing input, to do pay adjustments the model or training methodology. After designing model and reaching a model with an appropriate accurately according to input training and testing, if the answer models proper estimation to data, the model is ready for practical use. Otherwise should correct the design process [13-14].

Enhancing task SVM classification of data is based linear. The linear dividing data has tried to select the line that to be more reliable margin. In general, solve the equation to find optimal line for data by QP methods that methods are known in solving problem that is limited [15-16].

For a detailed study of the SVM algorithm: suppose, an optimal separating screen which is completely separate, with hyper plane with a maximum margin linear boundary exists. The training data is include N pair $(x_n, y_n),...,(x_2, y_2),(x_1, y_1)$, $x_i \in R^m$ *and* $y_i \in \{-1,1\}$,

Due to this we want profile pages to define a separator between two floors of 1 and -1, where the largest bond between two clouds parallel plates on each side of the cloud separator page, to

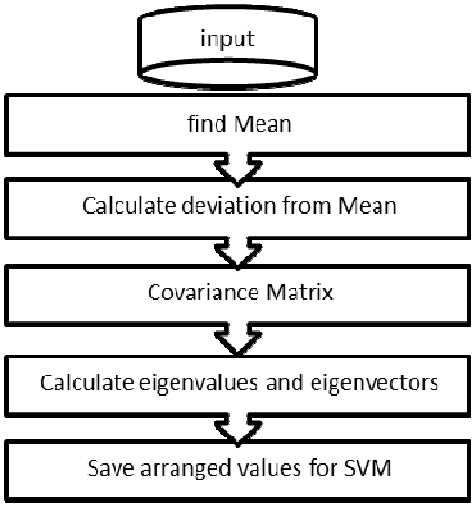

Figure2. PCA algorithm for Pre-processing

be there. For optimized separating Page the two classes are separated as with d with the nearest points of each class will have a maximum distance. Not only does this create a separator page to select a unique solution, but also with maximizing the bond between the two floors, shows a better performance in the separation of test data. In simple terms separator designed to extend the capabilities of a better whole. Then we discuss the optimization problem [17-18]:

$$\begin{aligned} &max \quad c \\ &w, w_0, \|w\| = 1 \end{aligned} \tag{1}$$

Where the constraint $i = 1,...,n$ and $y_i(x^T w + w_0) \geq C$. These adverbs are subject beyond ensuring a minimum distance of C in all parts of the boundary decision that w and w_0 are determined not to violate. For this context, we are looking for the largest C and related parameters that provide the conditions for us.

In fact did not possible implementation SVM, such that the line can be completely separated the data into distinct categories. In fact, data always have some flat of boundary Separator categories. This little flat is shown with the covariates $\xi = (\xi_1, \xi_2, ..., \xi_n)$. Classification error occurs when have ξ_i >1, by limiting $\sum_{i=1}^{n} \xi_i$ to value of *K* we obtain the optimization problem [16].

On the other hand, to resolve all needs and also satisfy the KKT conditions for this equation to equation (2) write.

$$\min_{w \cdot w_0} \frac{1}{2}\|w\|^2 \tag{2}$$

that for each i, $y_i(x^T w + w_0) \geq (1 - \xi_i)$ with condition $\xi_i \geq 0, \sum \xi_i \leq K$.

From this equation, it is well known that the points have been well side its class do not very important role in shaping the boundaries and this is a feature of this method [20-21].

The following equation is used to map the input space:

$$f(x) = \sum_{i=1}^{n} a_i y_i(\varphi(x), \varphi(x_i)) + w_0 \quad (3)$$

Moreover should have relation of kernel functions (the inner product in has converted space) an individual. Nuclear equation, with formula $k(x, x') = \sum_{j=1}^{m} \varphi_j(x)\varphi_j(x')$, to rewrite the formula (3) we use the following:

$$f(x) = \sum_{i=1}^{n} \hat{a}_i y_i(x, x_i) + \hat{w}_0 \quad (4)$$

The four core functions that are commonly used in SVM is, Linear function, polynomial function of degree d, the radial basis function (RBF) and MLP function (perception).[22-23] Steps in the algorithm in Fig2is shown. This flowchart symbolically are depicted the process performed on simulation algorithm based on support vector machines and analysis the main elements.

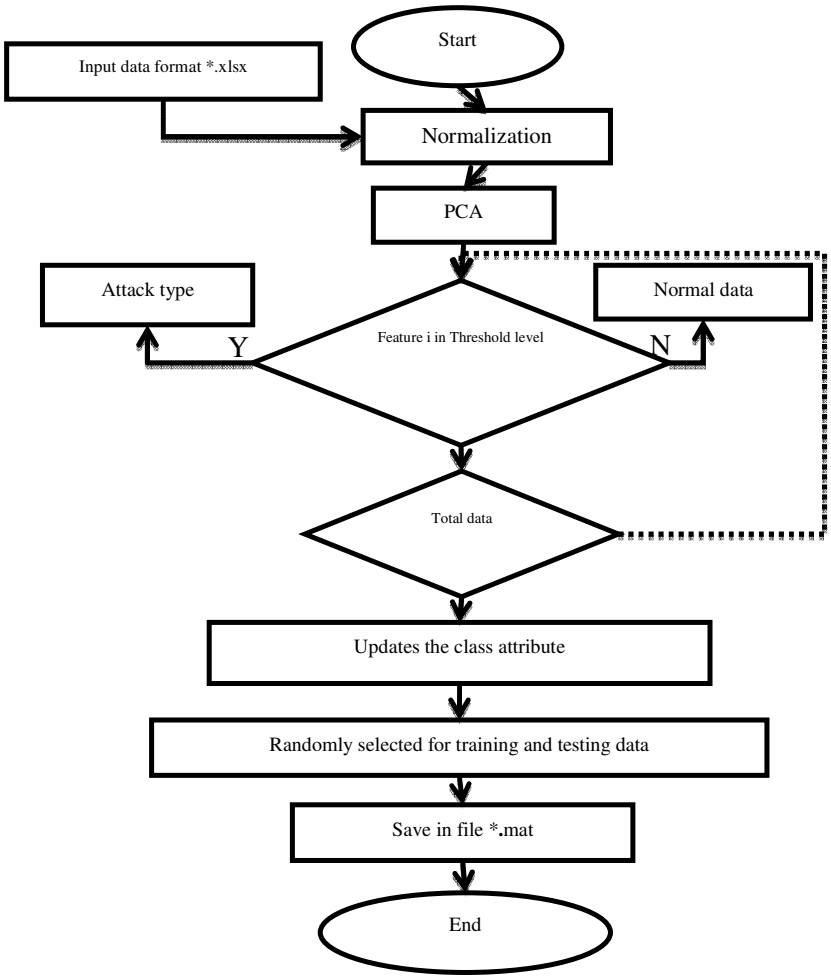

Figure 3. Flowchart of used algorithm for intrusion detection system

5. ANALYSIS OF RESULTS

In this simulation, presented algorithms is studied for total of 10% the initial data with different kernel and based on the dimension reduction method and the result is studied in term of response time, increase network efficiency, system error rate and sensitivity. KDD99 data set used in this simulation that main reason for using it the complete data set of all currently known attacks compared to other dataset used in the simulation experiments that have 41 attributes, which 21 kinds of abnormalities have in their place. This 21 anomaly in four total categories are named DOS, Prob, U2R and R2L.In simulation support vector machine algorithm method to analyze the main elements with help of Principal Component Analysis to reduce the number of features and increase system performance; Of 41 features used in KDD99 selected 17 features, for increase response time and system performance. In this experiment, the number of features is less than the response time will be faster.

The main reason for using PCA 52.7% improve response time and the increasing algorithm performance in intrusion detection from 99.40 to 99.84 and the error rate dropped to 26.6%.

In the simulations performed, the algorithm will be trained and then tested. During training, each group individual anomaly is compared with normative data but ultimately all abnormalities are placed a group. For training is used the radio labelled data, but during the test data are unlabelled. Ratio Data of each class to the total number of data in the data set are giveninTable1.

Table 1. Selected data distribution in data collection

R2L data	U2R data	Probe data	DOS data	Normal data	Total number
452.6123	21.45658	1591.43	158551.6	39398.33	200015

The simulation results of the PCA and support vector machine algorithm described in Table2.

Table 2. Numerical results for diagnostic tests

Algorithm is used	Correct Rate	Error Rate
SVM	99.4	0.6
PCA + SVM	99.84	0.16

The result of the simulation support vector machine Algorithm and impact of PCA on it with 41 features, 21 different types of abnormalities and different kernel is described below.

This simulations have 40, 000 Number of Observations, two Control Classes, one Target Classes, Inconclusive Rate 0,Classified Rate1 and Prevalence 0.8034.The results of RBF kernel when using PCA with different δ in Table 3 listed.

Table 3. Numerical results for different δ

δ	Correct Rate	Error Rate	Sensitivity
0.1	0.9698	0.0302	0.9625
0.5	0.9926	0.0074	0.9909
4.5	0.9973	0.0027	0.9968

Sigmoid kernel in case of using PCA is further tested for different α and β and the results are showninTable4. According to the table 4 can be found that, Change in the range of [α β] in manner that increases α and β decreases, Increases the error rate and increase the negative likelihood And therefore it be more possible to negative predictive value.

Table 4. Numerical results for different α and β

α	β	Correct Rate	Error Rate	Sensitivity
0.1	-0.6	0.9605	0.0395	0.9609
1.5	-0.6	0.9071	0.0929	0.9189
1.5	-1.6	0.8154	0.1846	0.9976
4.5	-0.6	0.8970	0.1030	0.9668

Result of the simulation with polynomial kernel and power3 can be seeninTable5.

Table 5. Numerical results for polynomial kernel with power 3

p	Correct Rate	Error Rate	Sensitivity
3	0.9771	0.0029	0.9991

Because the data in experiment are not consisted linearly and regular distribution, algorithm simulation with the PCA algorithm could not able to classify the data with a straight line .So we are unable to use of linearly kernel function in this simulation.

Evaluate the impact of using PCA algorithm for intrusion detection, the algorithm error will be have during the detection when using the PCA and not using it Figure 4 and 5 are shown. to be seen correct rate and error rate when principal component analysis is used for selected features in the kernel RBF and when principal component analysis is not used in quadratic kernel and linear kernel the best value have, according to compared correct rate, RBF kernel that chosen features with principal component analysis of the other kernel is better.

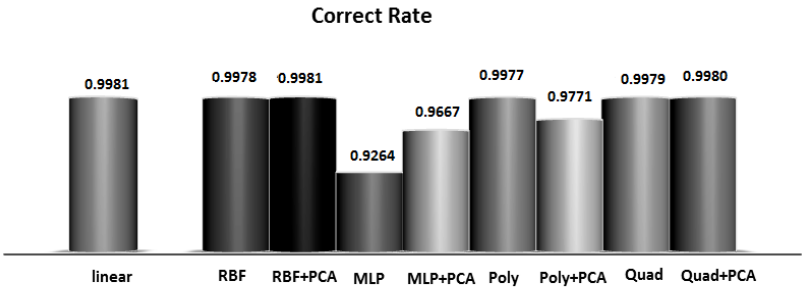

Figure 4. Numerical results for correct rate and impact of using PCA

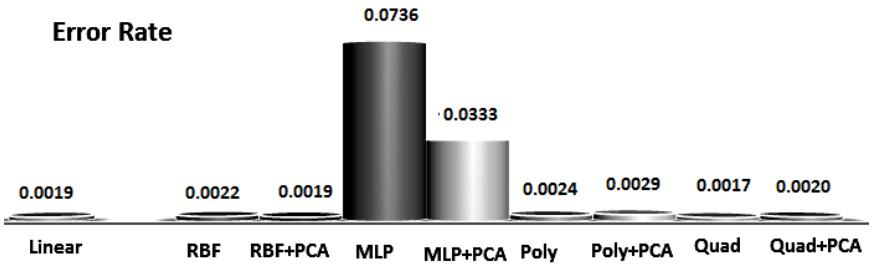

Figure 5. Numerical results for error rates and the impact of using PCA

6. CONCLUSION

Given the widespread use of smart metering devices in the context of information technology in the field of energy to manage user accounts that use networks for transferring data of readers' measurement devices causes an increasing topic for attackers. Systems must be designed to prevented manage data traffic over the network from attacks and sabotage activities in the field of information technology. In this paper, effect of different kernels of used SVM in the attacks classification algorithm i.e. intrusion detection systems exploiting support vector machine and PCA as pre-processing for separating the normal activity of network attacks is assessed. Because of the large number of features detected attacks and takes the vast amount of the computation, principal component analysis is used widely to reduce dimension. To explore effect of different kernel in intrusion detection system based on support vector machine and PCA, standard data KDD99 is applied in the algorithm and different kernel support vector machines have been evaluated. The results show that the correct detection rate and the rate of attack error detection when using principal component analysis in all cores radial, quadratic and linear of the lack of main components analysis have best value and total Radial Kernel accurate rate using the principal component analysis of all cores is better.

REFERENCES

[1] D. Dillona, J. Wheeldona, R. Chub, G. Choib, C. Loya, "Summary of EPRI's Engineering and Economic Studies of Post Combustion Capture Retrofit Applied at Various North American Host Sites" , Energy Procedia, vol. 37, pp. 2349–2358, 2013.

[2] Dillon et al, "An Engineering and Economic Assessment of Post-Combustion CO2 Capture Applied to FirstEnergy's Bay Shore Station Circulating Fluidized Bed Unit: Retrofit Study Report 5, EPRI Report 1019398. December 2011.

[3] http://www.iransg.com/fa/knowledge/articles

[4] Dillon et al, "An Engineering and Economic Assessment of Post-Combustion CO2 Capture applied to Nova Scotia Power's Coal-Fired Lingan Station: Retrofit Study: Report 3" EPRI Report 1019396. December 2011.

[5] AEP Smart Grid Demonstration Host- Site Overview Product ID 1020226.

[6] American Electric Power (AEP) Smart Grid Demonstration Host-Site Project Description Product ID 1020188.

[7] V.Das, V.Pathak, S.Sharma, Sreevathsan, M. Srikanth, G.Kumar, "Network Intrusion Detection System Based on Machine Learning Algorithms, "International Journal of Computer Science & Information Technology (IJCSIT), vol. 2, no. 6,PP. 138-151, 2010.

[8] M. Hasan, M. Nasser, B. Pal, S. Ahmad, "Intrusion Detection Using Combination of various Kernels Based Support Vector Machine," International Journal of Scientific & Engineering Research, vol. 4, no. 9, 2013 .

[9] Heba F. Eid, Ashraf Darwish, Aboul Ella Hassanien, and Ajith Abraham, " Principle Components Analysis and Support Vector Machine base Intrusion Detection System", IEEE 2010.

[10] S. Theodoridis, A. Pikrakis, K. Koutroumbas, and D. Cavouras, Introduction to Pattern Recognition with MATLAB, Pashalidis Pubs [In Greek]. 2010

[11] DaveDittrich, Network monitoring/Intrusion Detection Systems (IDS), University of Washington, Available Online At:

[12] L. I. Smith "A Tutorial on Principal Component Analysis." Available at: http://csnet.otago.ac.nz /cosc453/student_tutorials/principal_components.pdf. Accessed 2013-11-08.

[13] X.-Y. Wang and C.-Y. Cui, "A novel image watermarking scheme against desynchronization attacks by SVR revision," Journal of Visual Communication and Image Representation, vol. 19, pp. 334-342, 2008.

[14] A. Zainal, M. Aizaini Maarof and S. Shamsuddin, "Feature selection using rough set in intrusion detection", Tencon 2006, IEEE Region Conference, pp.1-4, 2006.

[15] L. Chun-hua, L. Zheng-ding and Z. Ke, "An image watermarking technique based on support vector regression", IEEE International Symposium on, Communications and Information Technology, vol. 1, pp. 183-186, 2005.

[16] T. Hastie, R.Tibshirani, and J.Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer, New York, 2001

[17] M. Tavallaee, E. Bagheri, W. Lu, and A.A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set", Proceedings of the Second IEEE international conference on Computational intelligence for security and defense applications, Ottawa, Ontario, Canada: IEEE Press, pp. 53-58, 2009

[18] S. Albayrak,F Amasyali., "Fuzzy c-Means Clustering on Medical Diagnostic Systems," International XII. Turkish Symposium on Artificial Intelligence and Neural Networks –TAINN, 2003.

[19] G. R. Zargar, P. Kabiri, "Selection of Effective Network Parameters in Attacks for Intrusion Detection, ICDM 2010, pp. 643-652, 2010.

[20] A. H. Sung, and S. Mukkamala, "The Feature Selection and Intrusion Detection Problems", Springer Verlag Lecture Notes Computer Science 3321, pp. 468-482, 2004.

[21] J. H. Friedman, "Multivariate Adaptive Regression Splines", Annals of Statistics 19, PP 1-67, 1991.

[22] H. G. Kayacık, A. N. Zincir-Heywood, and M. I. Heywood, "Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets," in Third Annual Conference on Privacy, Security and Trust , St. Andrews, New Brunswick, Canada, 2005.

[23]A. Iftikhar, B. Azween , A. Abdullah, M. Hussain: "Optimized intrusion detection mechanism using soft computing techniques," Telecommunication Systems, vol. 52, no. 4, pp. 2187-2195, 2013.