An Imperial learning of Data Mining Classification Algorithms in Intrusion Detection Dataset

Ms Pooja Bhoria[1], Dr. Kanwal Garg[2]

Abstract--- Classification is one of the important functionality of data mining. It performs its task by classifies the data into different categories using varieties of algorithms. This classification has wide range of applications in the field of intrusion detection in networking. It categorises network patterns as normal or attack to identify malicious activities occurring in the network. It enforces reliability to network users towards safe networking. In this paper, Researcher is analysing these classification algorithms which includes C4.5 decision tree, Naive bayes, Support vector machine (SVM), Self organised maps (SOM), Random forest, SVM Regression using NSL KDD 99 dataset using Orange canvas V2.6.1 data mining tool.

Keywords-- C4.5 Decision tree, Naive Bayes, Self organised machine (SOM), support vector machine (SVM), Random forest.

I. INTRODUCTION

Data mining is knowledge discovery process, which uses classification as one of the important functionality for extracting useful information from large volume of data [1],[2]. Classification includes wide variety of algorithms. These algorithms serves categorization of data into classes by creating a Model and then use that model to determine the class of newly arriving patterns [3],[4]. This categorization encapsulated in data mining has diverse applications in Intrusion detection system.

Intrusion detection system (IDS) identifies attack patterns occurring in the network and generates alarm when it encounters such malicious activities that violates security principles i.e. integrity, confidentiality and availability of stored data [1],[3],[5].

- **Ms Pooja Bhoria** presently pursuing M Tech in CSE in Department of Computer Science And Applications, Kurukshetra University, Kurukshetra, India. My area of research is data mining, in which I tried my best to use data mining for security applications. E Mail ID: poojabhoria22@gmail.com
- ***Dr Kanwal Garg*** *presently working as Assistant Professor in Department Of Computer Science And Applications, Kurukshetra University Kurukshetra, India. Owe the credit of more than 50 research papers published in international & national journals, conference & seminar. His area of expertise is Data Bases, Data Mining, & data warehousing. E Mail ID: gargkanwal@gmail.com*

This task of detecting intrusions is done by installing IDS at system nodes which continuously extract the incoming and outgoing network traffic patterns with the help of sensors [2],[3].Depending on the type of patterns stored in database of IDS, IDS is categorized into three categories i.e. 1) Misuse detection strategies: It is based on historical knowledge of attacks and stores all possible attack patterns in database. It poses good accuracy and low false alarm rate but it fails to deal with new attacks [5]. 2) Anomaly detection strategies: It stores the normal behaviour patterns of network and have high false alarm rate as deviation from patterns stored is considered as attack [6]. 3) Hybrid detection strategies: It stores normal behaviour patterns and attack patterns as well. It offers good accuracy and very low false rate. [6]

To detect malicious activities covered under these three categories, Data mining classification algorithms follows two phase process. In first phase, It creates a model of the patterns stored in database of IDS and then during the second phase, It classifies the pattern obtained (through sensors) from the network stream as normal or anomaly by using the model generated. If the pattern classifies as anomaly, it generates notification through alarms, e mails etc. Otherwise it remains silent [4].

To frame the objective stated before, this paper is divided into six sections. Section one comprises of introduction. Section two portrays IDS dataset description and tool information. Section three comprises of implementation of algorithms. Section four describes analysis and interpretation of classification algorithms. In section five, we finally conclude the paper and then section six comprises references.

II. IDS DATASET DESCRIPTION AND TOOL INFORMATION

NSL KDD dataset is offline network data based on KDD 99 dataset. It provides benchmark to the researchers to evaluate intrusion detection using offline data. This dataset has about 4,90,000 single connection records with no redundancy[9]. Along with removed redundancy, NSL KDD includes records to each difficulty level inversely proportional to number of records present [7],[13]. Each connection record has 41 attributes and one class attribute. Class attribute labels connection as normal or anomaly with exactly one specific attack types. Here, In this paper author is analysing 20% of the KDD training dataset with 10 folds cross validation and only determines the patterns as normal or anomaly. All the experiments are performed on Orange canvas version 2.6.1 data mining tool. It provides unified benchmark for researchers to analyse the learning model. Along with that, it also provides better user interface to users to get appropriate workflow schema.

III. IMPLEMENTATION OF CLASSIFICATION ALGORITHMS

This paper gives detailed analysis of varieties of data mining classification algorithms including C4.5 Decision tree, Naive bayes, SVM(support vector machine), SOM(self organised maps), random forest, SVM regression. Figure 1 shows the work flow schema in Orange canvas data mining tool as shown below. Among all the analysed classification algorithms, C4.5 decision tree provides better accuracy as it post-prune the tree after it is created that removes noise from the output tree generated but it suffers from a serious disadvantage, It generates unstable trees [10],[14],[12]. C4.5 deals much better with continuous as well as discrete attributes while naive bayes does not perform better with both type of attributes. To provide better classification from naive bayes, data must be pre processed first and then applied for classification. Also, Naive bayes classification algorithm is quite simple, robust and elegant that perform more efficiently with large databases as the graph produced is much smaller than C4.5 but it suffers from oversensitivity when number of attributes in dataset are strongly inter related to each other which effects the entire performance [11],[10].

SVM is eager learning algorithm which provides better generalisation capability than other classification algorithms for the data not classified properly but it suffers from low computational efficiency [10],[12]. It has much better capacity to deal with the outliers i.e. patterns that can't be classified by normal learning. This outlier detection is not provided by other machine learning algorithms discussed above. SVM regression is cross combination of SVM and regression which includes best features of both the classifiers. In terms of accuracy, it is quite similar to SVM but sensitivity is quite good than normal SVM [15]. Random forest is also eager learning algorithm that runs efficiently on large databases and provides better accuracy by generating many classification trees which polls to determine the class of patterns [16],[17]. But it consumes lot of resources like computational time, memory and CPU cycles. SOM maps multi dimensional nonlinear statistical data into two dimensional space. The main set back of this technique is that the number of output nodes is predefined and only the adjacent nodes are taken as neighbourhood which effects the performance of algorithm [19].

IV. ANALYSIS AND INTERPRETATION

In order to evaluate the performance of analysed classification algorithms, Researcher follows two phase process.

1) During first phase, NSL KDD dataset is pre-processed as it includes continuous, discrete, and symbolic attributes. These attributes can't be applied directly for classification. During pre -processing, all the symbolic attributes like protocol_type, service, flags etc are mapped to their corresponding integer/numeric values. For small integer value range attributes like duration, urgent, wrong_fragment etc, scaling is applied to them. And for large integer value range attributes like src_bytes, dst_bytes etc, scaling is applied again. And then all these attributes are mapped in range of (0.0, 1.0) accordingly.
2) During the second phase, Analysis work is performed by appling pre processed dataset to classification algorithms on Orange canvas data mining tool. This tool is installed on system having Intel core 2 duo processor 2.0 GHz processor and I GB of RAM [13]. During analysis, Researcher perform 10 fold cross validation on NSL KDD benchmark network intrusion detection dataset. Standard parameters such as accuracy, specificity, sensitivity, precision, recall, ROC curve area, TP Rate, FP Rate are used to estimate the performance of IDS. All the features are strongly co related to each other as each determines the measure of accuracy of classification algorithms. Table 1 tabulates the experimental results of algorithms.

Accuracy of algorithms specified in Table 1 determines how correctly an algorithm identifies normal and attack patterns [18]. Analysis shows that C4.5 provides best accuracy followed by random forest, SVM, SVM regression, naive bayes and SOM Maps. Accuracy can also be depicted by some other measures i.e. confusion matrix and ROC Curve. Confusion matrix gives detailed overview of how much instances are correctly classified and incorrectly identified by classification algorithm as shown in figure 2 [6]. These incorrectly identified instances results in high false alarm rate which is the combination of false positive rate and false negative rate. Higher will be the false alarm rate, worse will be the accuracy so it must be as much less as possible. Figure 2 describes that C4.5 decision has lowest false alarm

rate then tracked by random forest, SVM regession, SVM, Naive bayes and then SOM.

ROC curve as shown in figure 3 is plot between sensitivity and specificity. Sensitivity (Y axis parameter) for algorithm is defined as how correctly it identifies attack patterns [18]. For better illustration, It is also evaluated in table 1 which depicts that Random forest provides better results than tracked by C4.5, SVM regression, SVM, Naive bayes, and then SOM. Specificity i.e. X axis parameter determines how correctly it identified normal patterns from all incoming patterns. It is also evaluated in table 1 for better visualisation. It depicts that C4.5 provides better specificity followed by Random forest, SVM, SVM regression, SOM map, and Naive bayes. Accuracy is the combination of sensitivity and sensitivity. Both of them must be high for better classification into normal and attack categories. Its value of cross combination is depicted by ROC curve. Area covered by classifier in ROC curve (named as AUC in Table 1) determines classification capacity of algorithm, greater the area covered, better will be algorithm and better will be accuracy. Table 1 and figure 3 shows that Random forest covers the whole area of curve whereas others covers area in the order C4.5, SVM, SVM random, naive bayes, SOM maps.

There are some more measures that contribute to calculation of accuracy. It includes precision and recall. Precision may be defined as ratio of predicted positives/ negatives which are actually positive/ negative [4],[6]. It is determined by true alarm ratio (true positive/ (true positive + false positive)) and false alarm ratio (false positive/ (false positive + true positive)) [6],[14]. For better classification, true alarm ratio must be high and false alarm rate must be as much lower as possible [4]. In terms of precision, Analysis shows that C4.5 is most precise which is followed by random forest, SVM, SVM Regression, SOM and naive bayes. Another measure, Recall specified in table 1 may be defined as ratio of actual positive/negative which is predicted positive/negative and is given by TP rate and FP rate [6],[18]. True positive rate, also called sensitivity is given by (true positive/ (true positive + false negative)) must be high for better pattern detection and false positive rate, given by (false positive/ (true negative + false positive)) must be low as much possible for better intrusion detection [6][18]. Analysis shows that random forest provides best sensitivity or TP rate followed by C4.5 decision tree, SVM regression, SVM, naive bayes, SOM maps.

Away from Accuracy, performance of classification algorithms can also be measured by calibration curve. It analyses that whether the actual probability of classification reaches estimated probability. It may also be defined as the plot between estimated probability of detecting normal/attack patterns and actual probability of normal/attack patterns detected respectively. For all analysed algorithm plot is shown in figure 4. It depicts that diagonal curve shows perfect calibration which means that algorithm is total deviated towards pattern detecting capability. On comparing the calibration plot from figure 4, it concludes that SOM Map and C4.5 decision tree follows the diagonal curve and hence provides ideal calibration than other classification algorithms. Others classification shown gradual up and downs in their detection capacity.

V. CONCLUSION

In this paper, Researcher statistically analysed the NSL KDD dataset with 10 fold cross validation. The analysis concluded that each of the analysed algorithms performs better in their domains. C4.5 decision tree provide best accuracy and also shows better calibration towards pattern detection capacity. Naive bayes performs much better with large datasets having continuous attributes but fails drastically in case of discrete attributes. In case of co related attributes, naive bayes accuracy as well as calibration is worse. But by pre processing data set, it improves. Another analysed i.e. SVM provides best sensitivity and this sensitivity is improved further when regression is incorporated with SVM and random forest also offers very good accuracy than all analysed classification algorithms but not better than C4.5 decision tree. SOM map and C4.5 decision tree algorithm provides ideal calibration towards pattern detecting. Others analysed gradually posses up and down which determines they does not provides ideal calibration.

ACKNOWLEDGMENTS

Author like to thanks to thesis guide Dr. Kanwal Garg for his deep efforts and support towards the development of this research work. Also, Author would like to thanks Lincoln laboratory for providing such a valuable dataset for research in the field of intrusion detection.

REFERENCES

[1] Therodoros Lappas, Konstantinos pelechrinis, "Data Mining Techniques for (Network) Intrusion Detection Systems", Dated: 10-11-2012, UC RiverSide, Riverside CA 92521.

[2] Sherish Johri [2012], "Novel Method for Intrusion Detection using Data Mining", Dated 10-01-2013, International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), ISSN: 2277-128X, Volume 2, Issue 4.

[3]Kanwal Garg, Rshma Chawla [2011], "Detection of DDOS attacks using Data Mining", Dated 07-01-2013, International Journal of Computing and Business Research (IJCBR), ISSN(Online):2229-6166, Volume 2, Issue 1.

[4] Prabhjeet Kaur, Amit kumar Sharma, Sudesh Kumar Prajapat [2012], Dated 02-01-2013, "MADAM ID for Intrusion Detection Using Data Mining", International Journal of Research in IT and Management (IJRIM), ISSN 2231-4334, Volume 2, Issue 2.

[5]H. Patel & J. Sarvakar [2011], "Analysis of Data Mining Algorithm in Intrusion Detection", Dated 02-09-2012, International Journal of Emerging Technology and Advanced Engineering (IJETAE), ISSN 2250-2459, Volume 1, Issue 2, U.V.

[6] Radhika Goal, Anjali Sardana, & Ramesh C. Joshi[2011], "Parallel Misuse and Anomaly Detection Model", Dated 12-12-2012, International Journal of Network Security (IJNS), PP. 211-222, Volume 14, Number 4.

[7] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, Ali A. Ghorbani [2009], "A Detailed Analysis of KDD CUP 99 dataset", Dated 12-02-2013, Proceedings of IEEE Symposium on Computational Intelligence in Security and Defence Applications (CISDA).

[8] http://nsl.cs.unb.ca/NSL-KDD/ NSL KDD dataset

[9]J. McHugh [2000], "Testing intrusion detection systems: A Critique of the 1998 and 1999 Darpa Intrusion Detection System evaluations as performed by Lincoln laboratory," ACM Transactions on Information and System Security, pp. 262–294, vol. 3, Issue 4.

[10] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, QiangYang, Heroshi Mosada, Geoffrey J Melachlen, Augus Ng, Bing Lin, Philip S Yu, Zhi-hua Zhou, Michael Steinbach, David J Hand, Dan Steinberg [2008], " Survey paper on Top 10 Algorithms in Data Mining".

[11] Chotirat ANN Ratanmahatana, Dimitrios Gunopules, "Scaling up the Naive Bayasian Classifier using Decision Tree for Feature Selection", Dated 11-11-2012.

[12] J. Koshal, Monark Bag [2012], "Cascading of C4.5 Decision Tree and Support Vector Machine for Rule Based Intrusion Detection System", Dated 10-01-2013, International Journal of Computer Network and Information Security (IJCNIS), Pages 8-20.

[13] Mr. Manish Jain, Prof. Vineet Richariya [2012], "An Improved Technique based on Naive Bayesian for Attack Detection", Dated 04-12-2012, International Journal of Emerging Technology and Advanced Engineering (IJETAE), ISSN 2250-2459, Volume 2, Issue 1.

[14] S. Gunasakaran, C. Chandrasekaran [2011], "A Survey on Automobile Industries Using Data Mining Techniques", Dated 10-11-2012, International Journal of Science and Advanced Technology (IJSAT), ISSN: 2221-8386, Volume 1, Number 4.

[15] Govindarajan Muthukumarswamy [2011], "Network Intrusion Detection using Support Vector Regression", Dated 10-02-2013, International Journal of Artificial Intelligence and knowledge discovery (IJAIKD), ISSN 2231-2021, Volume 1, Issue 2.

[16] Madhuri Nallamothu, Mrs. D.N.V.L.S. Indira[2012], "Collaborative Filtering and Random Forest Classification Algorithm for PROBE Attack Detection in Network", Dated 4-03-2013, International Journal of Engineering Trends and Technology, ISSN 2231-5381, Volume 3, Issue 4.

[17] Jiong Zhang, M Zulkernine, A Haque[2008], "Random Forest based Network Intrusion Detection Systems", Dated 10-04-2012, IEEE, ISSN 1094-6977,Volume 38, Issue 5.

[18] M. Revathi, T. Ramesh, "Network Intrusion Detection System Using Reduced Dimensionality", Dated 2-12-2012, Indian Journal of Computer Science and Engineering (IJCSE), ISSN: 0976-5166, Volume 2, Number 1.

[19] A S Aneetha, Dr. S Bose [2012], " The Combined Approach for Neural Networks and Clustering Techniques", Dated 12-04-2013, Computer Science and Engineering: An International Journal (CSEIJ), Volume 2, Issue 4.

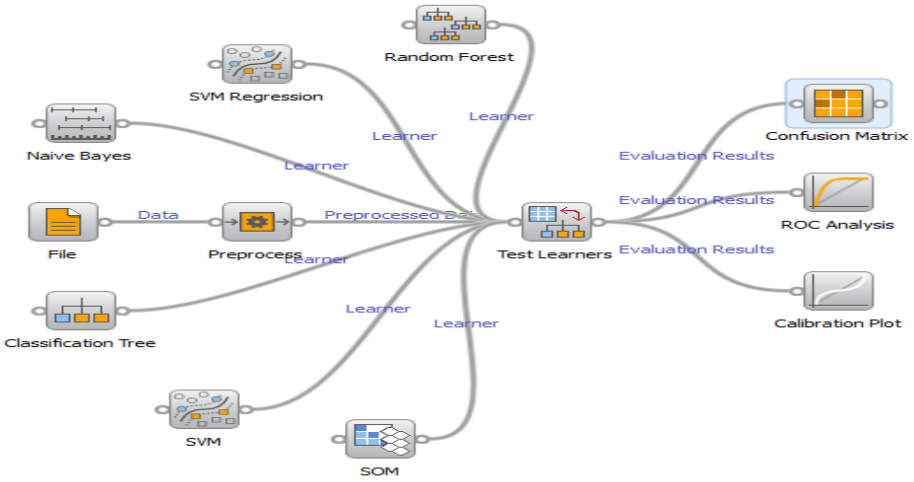

FIGURE 1: WORKFLOW IN ORANGE CANVAS DATA MINING TOOL

ALGORITHM	ACCURACY	SENSITIVITY	SPECIFICITY	AUC	PRECISION	RECALL
SOM	0.9022	0.8793	0.9284	0.9676	0.9336	0.8793
SVM	0.9738	0.9853	0.9601	0.9945	0.9658	0.9658
NAIVE BAYES	0.9230	0.9768	0.8613	0.9814	0.8897	0.9768
C4.5	0.9975	0.9983	0.9966	0.9990	0.9970	0.9983
RANDOM FOREST	0.9967	0.9995	0.9934	1.0000	0.9943	0.9995
SVM REGRESSION	0.9734	0.9876	0.9571	0.9923	0.9634	0.9876

TABLE 1: EXPERIMENTAL RESULTS OF ANALYSED CLASSIFICATION ALGORITHMS

	normal	anomaly	
normal	11826	1623	13449
anomaly	841	10902	11743
	12667	12525	25192

Self organised maps(SOM)

	normal	anomaly	
normal	13258	191	13449
anomaly	469	11274	11743
	13727	11465	25192

Support vector machine(SVM)

	normal	anomaly	
normal	13137	312	13449
anomaly	1629	10114	11743
	14766	10426	25192

Naive bayes

	normal	anomaly	
normal	13426	23	13449
anomaly	40	11703	11743
	13466	11726	25192

C4.5 decision tree

	normal	anomaly	
normal	13442	7	13449
anomaly	77	11666	11743
	13519	11673	25192

Random forest

	normal	anomaly	
normal	13282	167	13449
anomaly	504	11239	11743
	13786	11406	25192

SVM regression

FIGURE 2 : CONFUSION MATRICES OF CLASSIFICATION ALGORITHMS

.

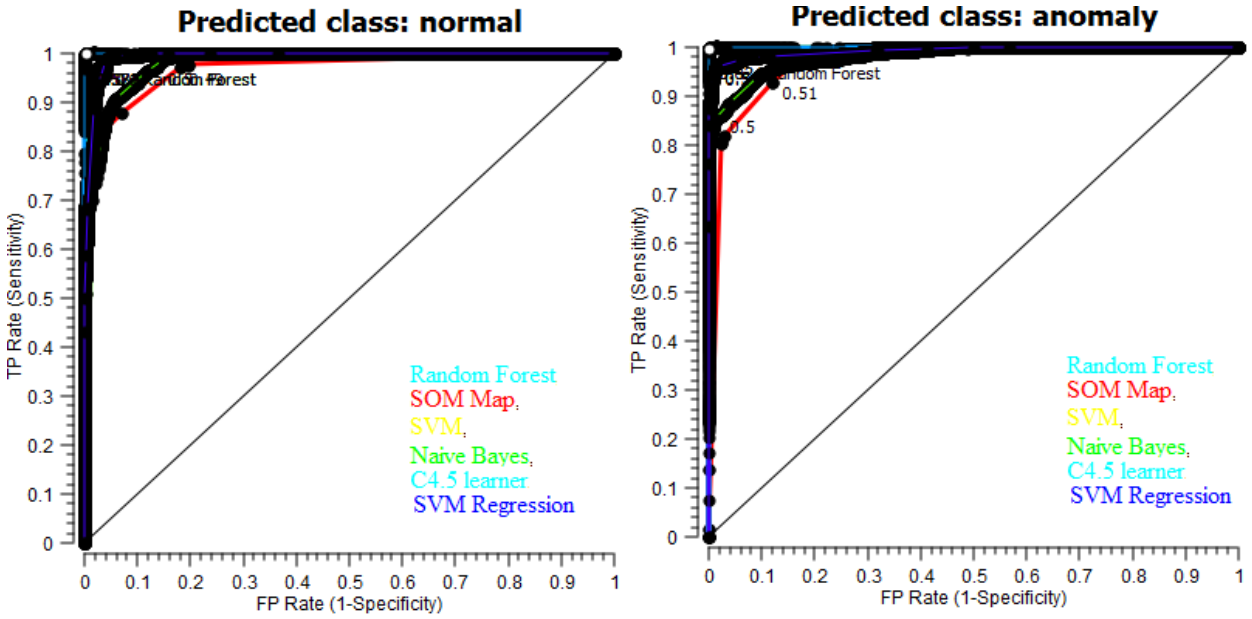

FIGURE 3 : ROC CURVE OF CLASSIFICATION ALGORITHMS

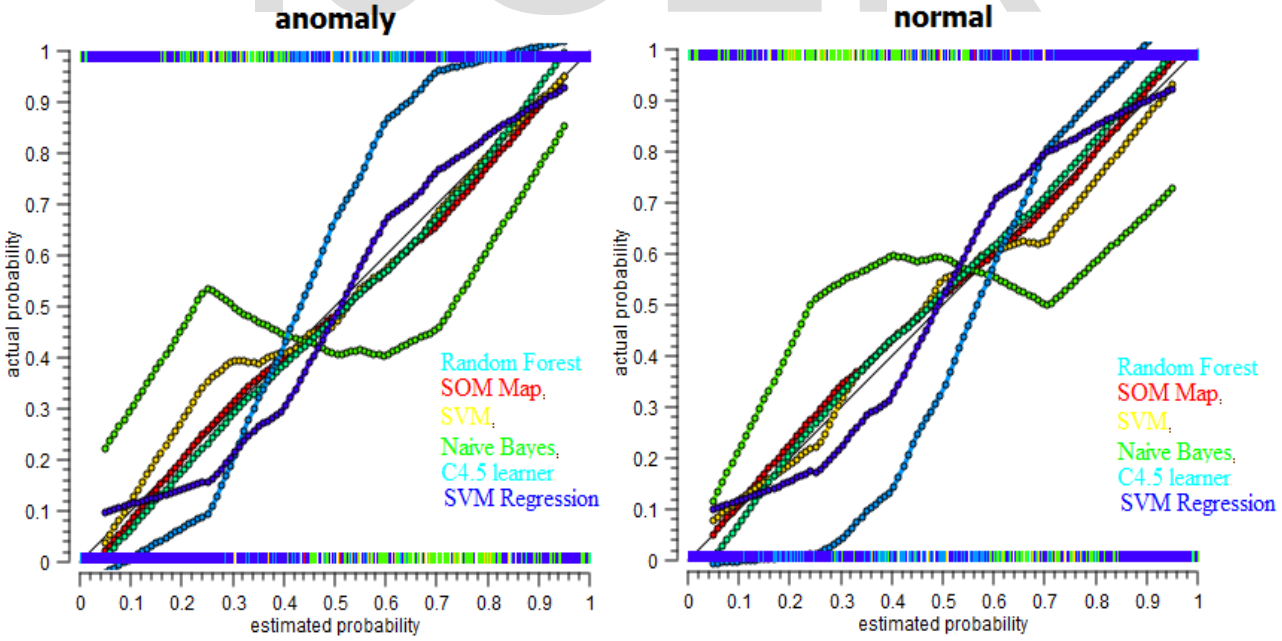

FIGURE 4: CALIBRATION PLOT OF CLASSIFICATION ALGORITHMS