Effective FAST and Fuzzy ARTMAP Performance to Detect Intrusion

Swati A Sonawale
PG Student
Department of computer Engineering
Dr.D.Y.Patil School of Engineering & Technology
Savitribai Phule Pune University
Pune,India

Roshani Ade
Assistant Professor
Department of computer Engineering
Dr.D.Y.Patil School of Engineering & Technology
Savitribai Phule Pune University
Pune,India

ABSTRACT
Great research work have been conducted towards Intrusion Detection Systems (IDSs) as well as feature selection. Feature selection applications have a great influence on decreasing development lead times and increasing product quality as well as proficiency. IDS guards a system from attack, misuse, and compromise. It can also screen network action. Network traffic observing and extent is progressively regarded as a key role for understanding and improving the performance and security of our cyber infrastructure. By using IDS attack can be detected in system as info is vital strength for every business. It can cause millions of harm within a few seconds. Security is important factor because reputation of business depends on it. So timely detection of intrusion is important so that preventive actions can be taken. IDS framework has been proposed by using fuzzy feature selection method with ARTMAP. It has been observed that the proposed framework gives better accuracy in less time as compared to methods in literature.

Keywords
Feature Selection, Intrusion Detection, Redundancy, Fuzzy ARTMAP.

1. INTRODUCTION
Feature selection has been an active research zone in pattern recognition, and data mining communities. The key notion of feature selection is to eliminate redundant features and select features which are needed [1].In order to deliver a clear image of the tradeoffs amongst the various ideas, feature selection has been framed as a multi-objective. As the dimensionality of a domain enlarges, the quantity of features n rises. Finding an ideal feature subset is intractable and problems related to feature selections have been tested to be NP-hard [2]. At this stage, it is important to define traditional feature selection procedure, which involves four elementary phases, i.e., subgroup, subset estimation, ending measure, plus validation. Subset generation is a search method that creates aspirant

Feature subsections for assessment grounded on a definite search strategy. Every candidate subset is assessed and matched with the prior superlative one according to a certain assessment. If the new subset turns to be well, it substitutes best one [3].This is repetitive way till a specified stopping state is fulfilled. Ranking of features picks the significance of any distinct feature, ignoring their probable communications [4].

2. LITERATURE SURVEY
Feature selection (FS) techniques have regularly been utilized as a principle approach to choose the important components. Unsupervised FS strategy, i.e., region furthermore, locality and similarity preserving embedding (LSPE) for highlight choice has been presented [5]. In particular, the closest neighbor diagram is firstly built to save the area structure of information focuses, and after that this territory structure is mapped to the remaking coefficients such that the closeness among these information focuses is saved. Besides, the sparsity inferred by the territory is likewise saved. At last, the low dimensional implanting of the inadequate recreation is assessed to best save the area and similarity [6].
Energy proficiency is a key issue in remote sensor systems where the vitality assets and battery capacity are extremely lacking. In this creator has presented another example acknowledgment based creation for diminishing the vitality utilization in remote sensor systems [7]. It includes a plan for calculation to rank and choose the sensors from the most vital to the slightest, and took after by a guileless Bayes classification. A proficient fuzzy classifier is rely on upon the capacity of highlight determination in view of a fluffy entropy strategy [8].It is utilized to assess the data of example dissemination in example space. With this data, we can isolate the example space into non covering choice areas for example arrangement. [9].At that point the choice areas don't cover, both the trouble and computational weight of the classifier is diminished and hence the preparation and arrangement time are short. In spite of the fact that the judgment territories are isolated into non covering subspaces, we can acquire great order execution in the meantime the choice regions can be appropriately distinguished by means of proposed fuzzy entropy technique[10].The element determination method decreases the dimensionality of an issue and in addition rejects clamor tainted, repetitive and inconsequential features. Today it is exceptionally fundamental need of procurement of an abnormal state security to watch very delicate and secret data. In Network Security Intrusion Detection System is a key innovation.[11].

These days' researchers have intrigued on interruption recognition framework utilizing Data mining strategies as a mischievous aptitude. IDS is a product or equipment gadget that arrangements with assaults by social occasion data from a variety of framework and system sources, then dissecting side effects. Various classification methods are discussed as follows [12].

2.1. The c4.5 tree-construction algorithm

The process builds a decision tree beginning from a training set T S, which consist of cases, or tuples in the database terminology. Each situation states values for a group of features and for a class. Every feature may have either one discrete or continuous values. Furthermore, the special value unknown is allowed, to denote unspecified values. The class may have only distinct values. We represent with C_1…… C_n Class the values of the class[13].

2.2. Decision Tree Classifier

Decision tree is a method of categorizing and forecasting data mining skill, belonging to inductive learning and supervised information mining technology. As decision tree is beneficial in fast creation and creating easy-to-implement if-then judgment rule, it has become the maximum widely useful practice amongst several classification approaches. Decision tree is one of the most widespread tools for classification and prediction. Construction of a decision tree is an effective scheme for grouping of data. This tree uses a top-down approach to form a test on each node.[14].

2.3. Naive Bayes Classifier

Naïve Bayes is a data mining method that displays success in classification Naïve Bayes is based on probability theory to find the best likely probable categorizations. According to Bayesian theorem, the probability of a set of data xt belonging to c is based on (1),

$$P(C|Xt) = \frac{P(C)P(Xt|C)}{P(Xt)} \quad \text{.....(1)}$$

Bayesian classifier computes provisional possibility of an example belonging to every class, and based on such provisional probability data, the instance is categorized as the class with the maximum provisional probability. In knowledge expression, it has the excellent interpretability same as decision tree, and is able to use previous data to build analysis model for classification [15].

Advantages/Disadvantages Of Naive Bayes

2.3.1 Advantages

1. Fast to train (single scan). Fast to classify.

2. Not sensitive to irrelevant features

3. Handles real and discrete data

4. handles streaming data well

2.3.2 Disadvantages

1. Assumes independence of features.

3. FAST ALGORITHIM

In our proposed FAST algorithm, it involves following steps:

1) Creation of the minimum spanning tree by using a weighted complete graph;

2) Distributing of the MST into a forest where each tree represents cluster.

3) Selection of representative features from the clusters.

INPUT: D (F1,F2…….Fm,C)

Θ=Threshold

OUTPUT:S=Set of selected feature subset

Step 1: For i=1 to m

1. T-relevance=SU(Fi ,C)
2. If T-relevance > Θ then
3. S=S U { Fi }
4. Create spanning tree by using prims algorithm
5. Make partition of tree to choose typical features.

Step 2: We get certain feature set.

4. SYSTEM ARCHITECTURE

Architecture of our system is as shown in below figure. Here we have taken KDD99 dataset as a input, then dataset is divided into 2 parts labeled and unlabeled dataset. On labeled dataset we are applying symmetric uncertainty with respect to each feature and we will find relevance of feature then we will generate spanning tree by using FAST algorithm of feature selection. On unlabeled dataset we are applying constraint selection algorithim.Therefore we are selecting only selected features. By using that features we can discover whether there is attack or not with the help of fuzzy ARTMAP classifier.

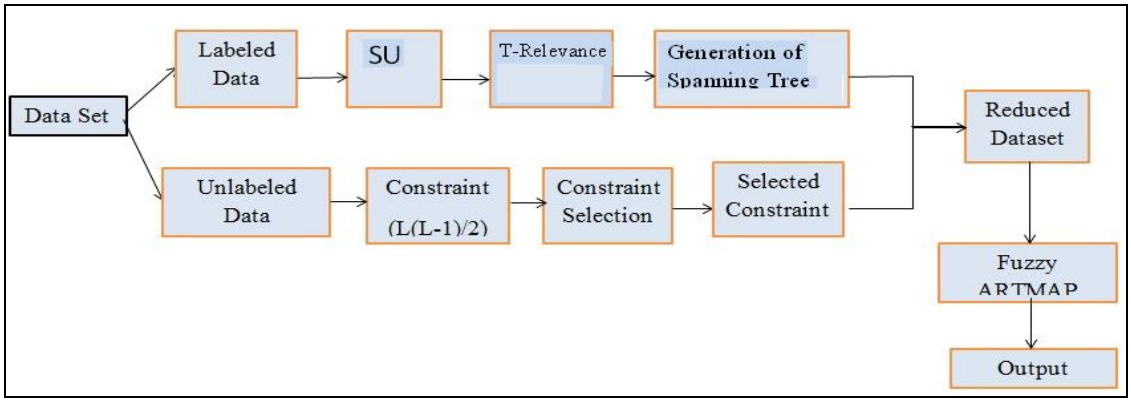

Figure 1. Proposed System Architecture

5. RESULTS

5.1. Comparison of ARTMAP with and without feature selection

Following figure shows comparison between ARTMAP with feature selection and ARTMAP Without feature selection in terms of classification accuracy

.

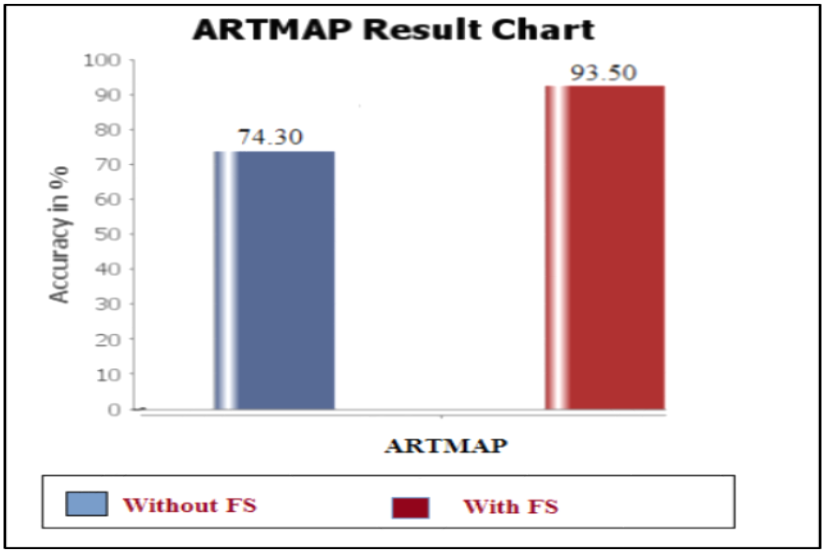

Figure 2: Comparison of ARTMAP with feature selection and without feature selection

5.2 Comparison of ARTMAP classifier with other classifier (ie. SVM)

Following figure shows comparison between ARTMAP with other classifier. Here ARTMAP is compared with SVM for accuracy. It has been observed that ARTMAP shows more accuracy as compare to SVM.

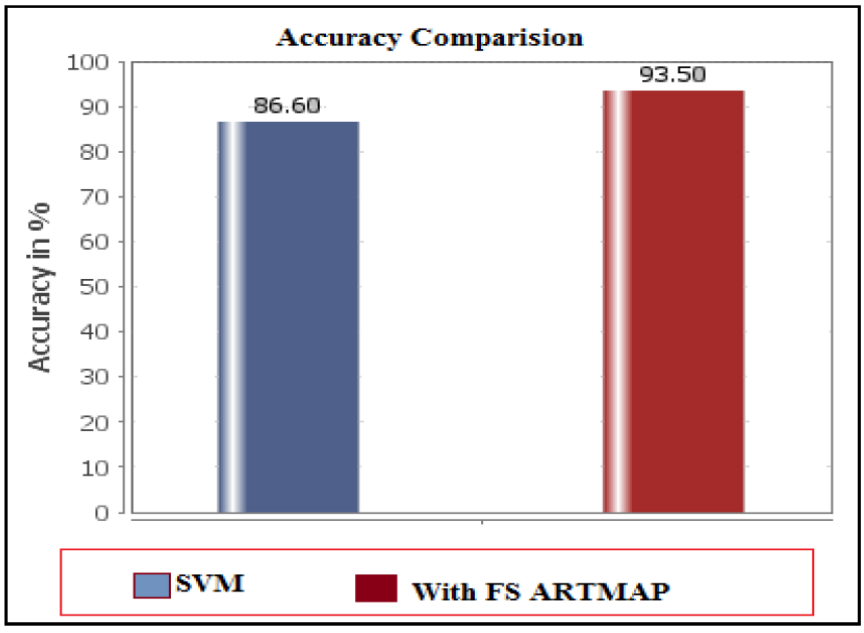

Figure 3 . Comparison of ARTMAP with other classifier(i e SVM)

5.3 Accuracy Of ARTMAP & NB Algorithm With FAST FeatureSelectionAlgorithm

To evaluate performance of proposed FAST algorithm is compared with 2 other classifier like NB, C4.

Sr.No	Classifier	Original Dataset	Reduced Dataset
1	ARTMAP	74.3%	93.5%
2	NB	90.1%	90.3%
3	C4.5	97.5%	97.7 %

Table 1: Performance of ARMAP, NB, C4.5 classifier with and without feature selection

6. CONCLUSION & FUTURE ENHANCEMENT

Thus we have proposed method of FAST algorithm with fuzzy ARTMAP which is used to reduce high dimensional data and within less time it will give accurate results which we are using to detect intrusion has occurred or not.

Data reduction technique is very useful because if we are getting same results within less time then why to process such huge amount of data and wasting processing capability.

We have compared the performance of FAST algorithm with naïve Bayes and c4.5.It has been observed that FAST algorithm gives more accuracy than any other classifier. In future alternative classifier can be chosen for classification so that it may increase performance.

7. REFERENCES

[1] A Fast clustering based feature subset selection algorithm for high dimensional data Qinbao Song, Jingjie Ni, and Guangtao Wang ieee transactions on knowledge and data engineering, VOL. 25, NO. 1, Jan 2013.

[2] Z. Zhao and H. Liu, Spectral Feature Selection for Data Mining, USA: Chapman and Hall-CRC, 2012.

[3] I.. Jolliffe, Principal Component Analysis, USA: Springer, 2002.

[4] X. He and P. Niyogi, "Locality preserving projections," in Proc. NIPS, 2004.

[5] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in Proc. NIPS, 2002.

[6] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, Mar. 2003.

[7] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," J. Mach. Learn. Res., vol. 5, Aug. 2004, pp. 845–889.

[8] M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of relief and relieff," Mach. Learn., vol.53, no. 1–2, pp. 23–69, 2003.

[9] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," J. Mach. Learn. Res., vol. 5, Oct. 2004, pp. 1205–1224.

[10] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in Proc. 24th Int. Conf. Mach. Learn., Corvallis, OR, USA, 2007.

[11] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in Proc. NIPS, Vancouver, Canada, 2005.

[12] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, "Feature selection via dependence maximization," J. Mach. Learn. Res., vol. 13, no. 1, Jan. 2012, pp. 1393–1434.

[13] Z. Zhao and H. Liu, "Semi-supervised feature selection via spectral analysis," in Proc. SIAM Int. Conf. Data Mining, Tempe, AZ, USA, 2007, pp. 641–646.

[14] D. Zhang, Z. Zhou, and S. Chen, "Semi - supervised Dimensionality reduction ," in Proc. SIAM Int. Conf. Data Mining, Pittsburgh, PA, USA, 2007.

[15] O. Chapelle, B. SchÄolkopf, and A. Zien, editors. Semi- Supervised Learning. MIT Press, C ambridge, 2006.

[16] S. BASU, M . BILENKO, AND R . MOONEY, A probabilistic framework for semi-supervised clustering, in KDD'04, Seattle, WA, 2004, pp. 59–68.

[17] U. Brefeld, T. G¨A RTNER, T. SCHEFFER, AND S. WROBEL, Efficient co-regularized least squares regression, in ICML'06, Pittsburgh, PA, 2006, pp. 137–144.

[18] K. WAGSTAFF, C. CARDIE, S. ROGERS, AND S. SCHROEDL, Constrained k-means clustering with background knowledge, in ICML'01, Williamstown, MA, 2001, pp. 577–584.

[19] T. ZHANG AND R. K. ANDO, Analysis of spectral kernel design based semi-supervised learning, in NIPS 18, MIT Press, Cambridge, MA, 2006, pp. 1601–1608.

[20] Z.-H. ZHOU AND M. LI, Semi-supervised learning with co-training, in IJCAI'05, Edinburgh, Scotland, 2005.

[21] X. ZHU, Semi-supervised learning literature survey, Tech. Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2006. http://www.cs.wisc.edu/»jerryzhu/pub/ssl survey.pdf.

[22] J. G. Dy and etal. Unsupervised feature selection applied to content -based retrieval of lung images. Transactions

on pattern Analysis and Machine Intelligence, 25(3):373-378, 2003.

[23] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle regression. Annals of Statistics, 32:407–49, 2004.

[24] G. Forman. An extensive empirical study of feature Selection metrics for text classification. Journal of Machine Learning Research, 3:1289–1305, 2003.

[25] Swati Sonawale and Roshani Ade "Review on intrusion Detection using fuzzy ARTMAP with feature selection Technique "International journal of science and research (IJSR), ISSN (Online):2319-7064, vol-3 ,Issue 11,Nov 2014.

[26] Swati Sonawale and Roshani Ade "Intrusion detection system-via fuzzy ARTMAP in addition with advance semi supervised feature selection" International journal of data mining and knowledge management process(IJDKP),vol. 5,no. 3,May 2015.

[27] Swati Sonawale and Roshani Ade " Dimensionality Reduction: An effective technique for feature selection", International journal of computer application:0975-8887, Vol.117.no 3,pp 18-23,May 2015.