# Feature Subset Evaluation and Classification using Naive Bayes Classifier

G. Keerthika
M.E Computer Science and Engineering
P.A College of Engineering and Technology, Pollachi. Tamil Nadu, INDIA

D. Saravana Priya
Assistant Professor Department of Information Technology
P.A College of Engineering and Technology, Pollachi. Tamil Nadu, INDIA

**Abstract – Feature Reduction is the reduction of features. Most of the intrusion detection approaches focused on feature selection issues such as irrelevancy, redundancy and length of detection process. These issues will degrade the performance of system. The performance of the system is improved by three feature selection methods involving correlation based feature selection, Gain Ratio and Information Gain. The threshold based Naive feature reduction algorithm is used to reduce the features. The reduced features are further classified by Naive Bayes classifier to produce best performance to design Intrusion Detection System.**

**Index Terms – Feature selection, classification, Feature reduction, Intrusion detection.**

## 1. INTRODUCTION

Feature Selection is the essential step in data mining. Individual Evaluation and Subset Evaluation are two major techniques in feature selection. Individual Evaluation means assigning weight to an individual feature. Subset Evaluation is construction of feature subset. The general criteria for feature selection methods are the classification accuracy and the class distribution. The classification accuracy does not significantly decrease and the resulting class distribution, given only the values for selected features [22]. Feature Selection can support many applications, it include the problems involving high dimensional data [21].

Figure. 1 describes feature selection steps. The four key steps in feature selection are

- Subset generation
- Subset Evaluation
- Stopping criteria
- Result validation

The feature selection is used to select relevant features by removing irrelevant and redundant features to improve the performance and to speed up the learning process.

The training data arrive in sequential manner from real time application such as Intrusion Detection System, making it difficult to develop a regular batch feature selection. To avoid this limitation, online feature selection problem can be overcome by online learning techniques. The main goal of online feature selection is to deploy online classifiers for classification. Online feature selection is essential when a real time application has to deal with high dimensionality training data.

Feature reduction is needed to reduce the number of features that are required to find the attacks [16]. Feature reduction techniques involving correlation based feature selection, Gain Ratio and Information Gain is used to reduce the features. The reduced features will be classified by Naive Bayes classifier. We proposed the Naive Feature Reduction algorithm to improve the performance level and accuracy.

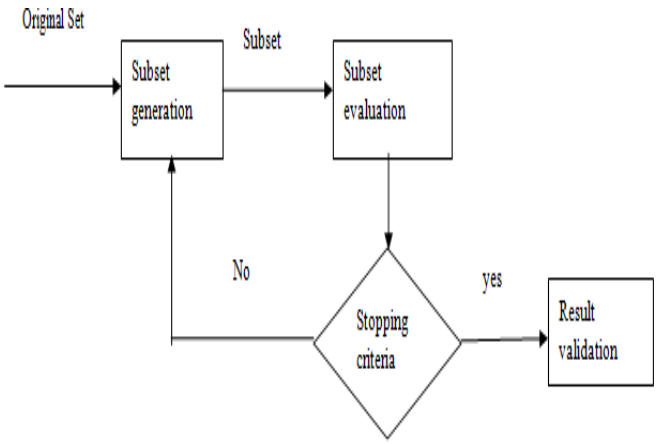


Figure. 1 The general procedure for feature selection

## 2. RELATED WORK

Feature selection can improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. The problem of Online

Feature Selection is aiming to resolve the feature selection problem in an online fashion by effectively exploring online learning techniques. In particular, two kinds of Online Feature Selection tasks are addressed in two different settings, learning with full inputs of all the dimensions or attributes and learning with partial inputs of the attributes [21].

Lei Yu and Huan Liu proposed a fast filter method to identify relevant features. Feature selection is preprocessing step to machine learning and effective in selecting relevant data by removing irrelevant data, increasing accuracy, and increasing result. However, the increase in dimensionality of data poses a severe challenge to many existing feature selection methods with respect to efficiency and effectiveness. A fast filter method can identify relevant features as well as redundancy among relevant features without pair-wise correlation analysis. The efficiency and effectiveness of the method is demonstrated through extensive comparisons with other methods using real-world data of high dimensionality [19].

Giovanni Cavallanti proposed shifting Perceptron algorithm achieves the best known shifting bounds on using an unlimited budget. It obtains efficient memory bounded algorithms with good theoretical shifting performance. Simple changes to the standard Perceptron algorithm suffice to obtain efficient memory bounded algorithms with good theoretical shifting performance. The randomized algorithms exhibited a substantially smooth convergence to the error rates of their non-budget counterparts. A shifting Perceptron algorithm achieving the best known shifting bounds by using an unlimited budget is introduced and analyzed.

Rajdev Tiwari and Manu Pratap Singh developed a correlation based feature selection using genetic algorithm. Integrating data sources is referred to as the task of developing a common schema. It is also data transformation solutions for a number of data sources. The size of the data should fit to datawarehouse. The features are reduced using Attribute subset selection and correlation analysis and it detects the unwanted features [28].

Jasmina novakovic, perica strbac, Dusan bulatovic implemented a comparison between several feature ranking methods. This method is implemented on datasets. The six ranking methods is divided into two types. They are statistical and entropy based method. The supervised learning algorithms are used to build models. Naive Byes, C4.5 are used for classification. Based on the ranking methods, the classification accuracy is obtained. In this work, ranking methods with different supervised learning algorithms give different results for balanced accuracy [30].

Various researchers have used Genetic Algorithm for optimization. Genetic algorithm is used to search the optimal technique for attribute selection. It validates method for optimal feature subset selection and it is used as optimal search tool for selecting features.

Huan Liu implemented the feature selection and dimensionality reduction. Data processing and storage has expand capabilities in production, communication and research. Feature selection is an important technique for feature reduction. It has been used to overcome the problem in processing high dimensional data. The advantages of this system are building simpler models, improving performance level and understand data [19].

Nicol `o Cesa-Bianchi proposed an efficient algorithm for learning linear predictors that actively samples the attributes of each training instance. The Perceptron algorithm performs well on online classification tasks. A common difficulty encountered during implementing kernel based online algorithm is the amount of memory required to store the online hypothesis and may grow unboundedly. It is a kernel based online learning algorithm with a fixed memory budget. It has the capability of bridging the small gap between Upper bound and budget constraint remains an open problem.

E.N. Lutu proposed a naive bayes classifier for classification. Stream mining is the process of mining a continuous, ordered sequence of data items in real time. Naïve Bayes classification is one of the popular classification methods for stream mining because it is an incremental classification method. The performance of the Naive bayes classifier improves by eliminating irrelevant features from the modeling process. It identifies efficient computational methods for selecting relevant features for Naive Bayes classification based on the sliding window method of stream mining [27].

Jialei Wang proposed an Online Feature Selection algorithm. Online learning was implemented with two methods i) Learning with full input ii) Learning with partial input. Learning with full input allows the learner to access all features [16]. Learning with partial input allows the learner to access only small number of features. Feature Selection algorithm uses binary classifier and it produces efficient and scalable result when applied to real world applications. Online feature selection algorithm was used to select relevant features and online classifiers were developed to classify several feature subsets.

Filter methods and wrapper methods are used for feature selection and feature reduction. Filter methods select important features by calculating correlation between feature subsets. Wrapper methods select important features subsets by using some predetermined algorithm.

## 3. FEATURE REDUCTION

Feature reduction is the important technique in data mining. Feature reduction is reduction of features and it leads to better understanding of prediction model.

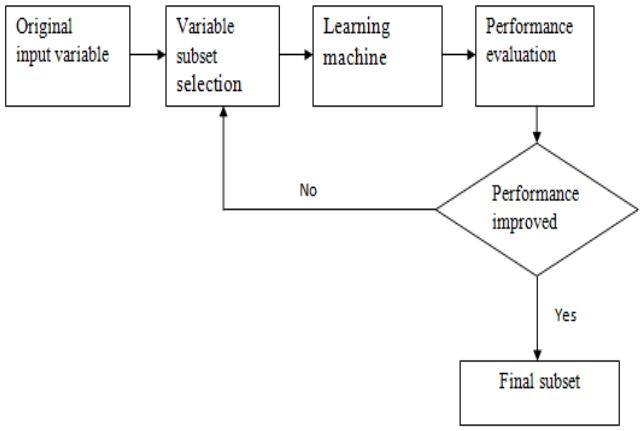


Figure. 2 Feature Reduction

There are two approaches in feature reduction. A wrapper approach evaluates features by using the learning algorithm. The filter approach evaluates features by understanding the general characteristic of data. The wrapper method produces better result but works slowly than a filter method. There are three feature reduction techniques including correlation based feature selection, Information gain and Gain ratio [15].

### 3.1. Correlation Based Feature Selection

Features are evaluated and ranked as a feature subset rather than the individual feature. It can select the set of features that are highly correlated but with low inter correlation. Correlation based feature selection calculates correlation between two features and also find the correlation between feature and class from the training and testing data.

$$M_k = \frac{s\overline{r_{cf}}}{\sqrt{s + s(s-1)\overline{r_{ff}}}} \quad (1)$$

Where $M_k$ is the correlation between the summed feature subset, s is the number of subset, $r_{cf}^{-}$ is the correlation between the class and features, $r_{ff}^{-}$ is correlation between feature subsets ([28], [23], [19]).

### 3.2. Information Gain

The information gain is measuring for each feature subset with respect to class and the selected features will be evaluated by Information gain [30]. Let B be set contains b data samples with n different classes. The dataset contains both training data and test data. The data is randomly divided into feature subsets. One of the subset is used as test data ant remaining subset is used as training data. The number of iteration increases based on the number of feature subsets [23].

The information gain ratio is calculated as

$$\text{Gain (F)} = I(b_1, b_2, \ldots, b_n) - E(F) \quad (2)$$

$I(b_1, b2, \ldots, b_n)$ is expected information and it is calculated to classify the samples. The information gain measures the amount of information in the features.The expected information is calculated as

$$I(b_1, b_2, \ldots, b_n) = -\sum_{i=1}^{n} \frac{b_i}{b} \log_2 \frac{b_i}{b} \quad (3)$$

Where b is data samples and $b_i$ is feature subset class.
Let F be features. The set of v distinct feature values be $\{f_1, f_2, \ldots f_v\}$. The training set is divided into v subsets. It can be denoted as $\{c_1, c_2, \ldots, c_v\}$. The expectation of F is calculated as

$$E(F) = \sum_{j=1}^{v} \frac{b_{ij}}{b} + \cdots + \frac{b_{nj}}{b} \text{x } I(b_{ij}, \ldots, b_{nj}) \quad (4)$$

Where $b_i$ is the subset, $b_i$ has the value $f_i$ for feature F and $b_i$ contains $b_{ij}$ samples of class i.

### 3.3. Gain ratio

The errors occurred during classification are training data noise, bias and variance. Gain ratio is an extension of information gain and it attempts to overcome the error involving bias. It applies normalization to information gain and it is defined as

$$\text{splitinfo}_m(b) = -\sum_{i=1}^{v} \left(\frac{|b_i|}{|b|}\right) \log_2 \left(\frac{|s_i|}{|s|}\right) \quad (5)$$

The above equation denotes the splitting of training data into v samples. Each sample is used as test data and training data. One feature subset is used as testing data and other feature subsets are considered as training data. After finding information gain, it applies normalization to information gain. By using split information, the gain ratio will be calculated. The gain ratio is

$$\text{Gain ratio (F)} = \text{Gain (F)}/\text{splitinfo}_F\,(k) \quad (6)$$

### 3.4. *Naive* bayes classifier

A naive bayes classifier is based on bayes theorem and it can achieve good performance in classification task. The

conditional probabilities is $P(x_j|c_i)$ and prior probabilities is $P(c_i)$. $P(c_i)$ are calculated by counting the training sample and then dividing the count result by training set size. A naive bayes classifier is defined as

$$F_i(x) = P(x_j|c_i)P(c_i) \quad (7)$$

Where $X = (x_1,x_2,\ldots x_N)$ represents feature vector, $C_j$ represents class labels, j=1, 2,…,N [25], [26].

The naive bayes model is a simplified Bayesian probability model. The naive Bayes classifier assumes that the features are independent and operates on it. The probability of one attribute will not affect the probability of other attribute. It works well on classification but it states that the error occurs due to three factors including bias, variance and training data noise. Training data noise can reduce by selecting good training data. The impact of bias is large and the impact of variance is small during grouping of training data ([9], [27]).

The naive bayes classifier will be evaluated on KDDCUP99 dataset to detect the attacks. The four categories of attack are Denial of Service, Probe, and Remote to Local and User to Root. The feature reduction is evaluated on feature subsets using feature selection methods involving Correlation based feature selection, Information Gain and Gain Ratio and the reduced datasets are further classified using Nave Bayes classifier. It classifies the feature subsets based on the number of features reduced.

### 3.5. Naive feature reduction method

The three main performance criteria are the classification accuracy, true positive rate, false positive rate. These performance criteria are used to find the vitality of feature. The searching technique is used to identify importance of features. The accuracy is measured by removing each feature one at a time. The process of feature selection continues until the classifier accuracy match with the original result [2].

The leave-one-out method is used to remove one feature from the original dataset, evaluate the experiment and then compare the new result with the original result. Based on the performance level, the importance of features is identified. On deleting a feature, decrease in performance will indicate the feature is important and increase in performance will indicate the feature is unimportant and no changes found in performance will indicate the feature is less important.

ALGORITHM

**Input**
F = Full set of features
Ac = accuracy of classifier
Avg_tpr = average TPR
//NFR

**Begin**

**Initialize:** set = {F}
For each feature {f} form
(1) T = set{f}
(2) Invoke Naive Bayes classifier
(3) If CA >= ac and RMSE <= err and A_TPR >= avg_tpr then
Set = Set-{f}
F = Set

**End**
**Output**
Accuracy
TPR value
FPR value

## 4. EXPERIMENTAL RESULT

KDDCUP99 dataset contains 41 features and each instance in features is named as normal or attack. In this dataset, there are 4,94,021 instances in which 92,278 are named as normal and 3,96,744 are named as attacks. There 22 types of attacks and these attacks are classified in four categories. They are Denial of Service, Probe, and Remote to Local and User to Root. Feature reduction is implemented with Java Netbeans.

Table 1.Types of Attacks

| DOS | Probe | Remote to Local | User to Remote |
|---|---|---|---|
| Back | Ipsweep | ftp_write | Loadmodule |
| Land | Nmap | guess | Rootkit |
| Neptune | Ports | passwd | Perl |
| Pod | Weep | imap | Normal |
| Smurf | Satan | multihop | bufferoverflow |
| Teardrop | Normal | phf | |
| Normal | | spy | |
| | | warzclient | |
| | | Normal | |

### 4.1. Performance Evaluation

Confusion Matrix is used to evaluate the performance of classifier. True positive means the actual class of features is positive but the classifier predicts the class correctly as positive. False positive means the actual class of features is negative but the classifier predicts the class incorrectly as

positive. True negative means the actual class of features is negative but the classifier predicts the class as negative.

True Positive Rate, TPR = TP /(TP+FN) (8)

False Positive Rate, FPR = FP/ (TN+FP) (9)

Classifier accuracy, CA = (TP+TN)/(TP+FN+FP+TN) (10)

| Actual class | Predicted class | |
|---|---|---|
| | Normal | Attack |
| Normal | TP | FN |
| Attack | FP | TN |

Table 2. Confusion matrix

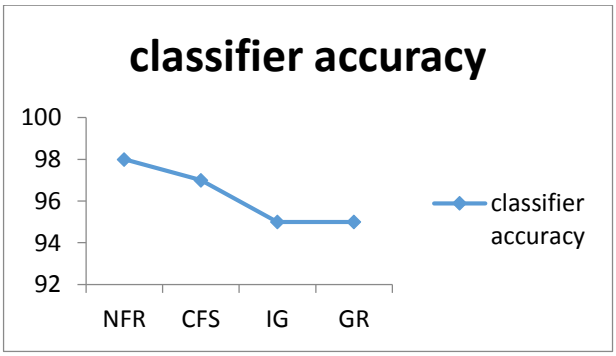


Figure. 3. Classifier accuracy

Figure 3 shows classification accuracy of Naive feature reduction, correlation based feature selection, Gain ratio and Information gain in percentage.

## 5. CONCLUSION

In this paper, we examined various techniques of feature selection. We can use Feature selection methods involving Correlation based feature selection, Information Gain, Gain ratio and Naive feature reduction to reduce the features. In future, we will customize NFR to improve the results for intrusion with reduced complexity and overheads and compare the efficiency rates with previous methods.

## REFERENCES

[1] Bekkerman R, "Distributional Word Clusters vs. Words for Text Categorization," Journal of Machine Learning Research, vol 3, pp 1183–1208, 2003.
[2] Bennett K. P, "Dimensionality Reduction via Sparse Support Vector Machines," Journal of Machine Learning Research, vol 3, pp 1229–1243, 2003.
[3] Cavallanti G. and Cesa-Bianchi N, "Tracking The Best Hyperplane with A Simple Budget Perceptron," Machine Learning, pp 143–167, 2007.
[4] Chan A. B, "Direct Convex Relaxations of Sparse Svm," In ICML, pp 145–153, 2007.
[5] Crammer K, "Singer Online Passive Aggressive Algorithms," J. Mach. Learn. Res. (JMLR), vol 7, pp 551–585, 2006.
[6] Crammer K, "Convex Confidence Weighted Learning," In NIPS, pp 345–352, 2008.
[7] Crammer K, "Adaptive Regularization of Weight Vectors," In NIPS, pp 414–422, 2009.
[8] Dash M. and Gopalkrishnan V, "Distance Based Feature Selection for Clustering Microarray Data," In DASFAA, pp 512–519, 2008.
[9] Dash M. and Liu H, "Feature Selection for Classification," Intell. Data Anal., vol 1, pp 131–156, 1997.
[10] Dekel .O, "The Forgetron: A kernel-based Perceptron on a Budget," SIAM J. Comput, vol 37, pp 1342–1372, 2008.
[11] Ding C. H. Q. and Peng H, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," J. Bioinformatics and Computational Biology, vol 3, pp 185–206, 2005.
[12] Donoho D. "Compressed Sensing," IEEE Transaction on Information Theory, vol 52, pp 1289 – 1306, 2006.
[13] Dredze M, "Confidence Weighted Linear Classification," In ICML, pp 264–271, 2008.
[14] Duchi J. and Singer Y, "Efficient Online and Batch Learning using Forward Backward Splitting," J. Mach. Learn. Res., vol 10, pp 2899–2934, 2009.
[15] I.H.Witten, E.Frank, M.A. Hall, "Data Mining Practical Machine Leanrning Tools & Techniques," Third edition, Pub. – Morgan kouffman.
[16] Steven C. H, "Online Feature Selection and Its Application," IEEE Transaction On Knowledge, vol 26, No 3, pp 698-711, 2014.
[17] Ranjit Abraham, Jay B. Simha and S. Sitharama Iyengar, "Effective Discretization and Hybrid Feature Selection using Naïve Bayesian classifier for Medical datamining," ISSN 0974-1259 Vol.5, No.2, pp. 116–129, 2009.
[18] M. Dash and H.Liu, "Feature Selection for Classification," Intell.Data Anal., 1(1-4):131–156, 1997.
[19] Lei Yu leiyu and Huan Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," ICML, 2003.
[20] Lei Yu and Huan Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," Journal of Machine Learning Research, pp 1205–1224, 2004.
[21] Vipin Kumar and Sonajharia Minz, "Feature Selection: A literature Review," Smart Computing Review, vol. 4, no. 3, 2014.
[22] Liu H, Setiono R, Motoda H, Zhao Z, "Feature Selection: An Ever Evolving Frontier in Data Mining," JMLR: Workshop and Conference Proceedings 10: 4-13 The Fourth Workshop on Feature Selection in Data Mining.
[23] Mark A. Hall, "Correlation-based Feature Selection for Machine Learning," Dept of Computer Science, University of Waikato, http://www.cs.waikato.ac.nz/~mhall/thesis.pdf.
[24] j.Han, M Kamber, "Data mining : Concepts and Techniques," San Francisco, Morgan Kauffmann Publishers(2001).
[25] Wafa' S.Al-Sharafat, and Reyadh Naoum, "Development of Genetic-based Machine Learning for Network Intrusion Detection," World Academy of Science, Engineering and Technology 55, 2009

[26] Ms.Nivedita Naidu, Dr.R.V.Dharaskar, "An Effective Approach to Network Intrusion Detection System using Genetic Algorithm," International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 2, 2010.
[27] Patricia E.N. Lutu, "Fast Feature Selection for Naive Bayes Classification in Data Stream Mining," Proceedings of the World Congress on engineering, Vol III, WCE 2013.
[28] Rajdev Tiwari, Manu Pratap Singh, "Correlation-based Attribute Selection using Genetic Algorithm," International Journal of Computer Applications, pp (0975 – 8887), Volume 4– No.8, August 2010.
[29] Nicol`o Cesa-Bianchi, Shai Shalev-Shwartz, Ohad Shamir, "Efficient Learning with Partially Observed Attributes," Journal of Machine Learning Research, pp 2857-2878, 2011.
[30] Jasmina novakovic, Perica strbac, Dusan bulatovic, "Toward Optimal Feature Selection using Ranking Methods and Classification Algorithms," pp 119-135, 2011