ISSN 2393-8021

International Advanced Research Journal in Science, Engineering and Technology
Vol. 1, Issue 2, October 2014

Mutual information based feature selection techniques for intrusion detection

Gulshan Kumar

Shaheed Bhagat Singh State Technical Campus, Ferozepur, Punjab, India

Abstract:A huge amount of high dimensional audit data is the major problem for accurate & quick detection of the intrusions. The audit data may contains some irrelevant & redundant features. Processing of these features by an IDS may increase the computational overhead, decrease the overall accuracy, and delay the process of intrusion detection. Therefore, for accurate & quick intrusion detection, the audit data may be reduced by selecting the most relevant and non-redundant features.In this paper, we explored various feature selection techniques especially mutual information (MI) based filter feature selection techniques. An updated review of the important techniques in literature is presented. The review will help the better understanding of different directions in which research has been done in the field of feature selection. The findings of this paper provide useful insights into literature and are beneficial for those who are interested in applications of MI based feature selection techniques to IDS and related fields. The review also provides the future directions of the research in this area.

Key words: Feature reduction Data Breaches Feature selection Intrusions Intrusion detection Network Security Security Threats

1 INTRODUCTION

Due to high availability of sensors, high processing speed and low cost storage devices, many applications of different fields produce data of high dimensions for analysis. In theory, higher dimensions of the data improve the classification accuracy of the algorithm. But, practically, it is not true. All the features of the data are not important to understand it. However, the higher dimensions of the data suffer from difficulty called curse of dimensionality [4]. The difficulty of analyzing the high dimensional data can be tackled in two ways. One way is to develop the technique that generates models which are able to analyze the high dimensional data efficiently. Another way is to reduce the dimensions of the data to process without loss of significant information. Feature combination transforms the features either linearly or non-linearly. Major techniques employed for feature combination are Principal Component Analysis (PCA), Independent Component Analysis, Linear Discriminant Analysis, etc. Feature selection or reduction keeps the original features as such and select subset of the features that predicts the target class variable with maximum classification accuracy [12]. In this work, the term feature selection, feature reduction or dimension reduction ia used interchangeably. The reduction in dimensions helps to improve the following:

1. Efficiency in terms of Measurement costs, Storage costs & Computation costs;

2. Classification performance;

3. Ease of interpretation/modeling.

The aim of feature selection techniques has many folds. It avoids the over fitting problem, improves the performance of classification models, develops fast and cost effective models, facilitates data visualization & data understanding, reduces the measurement and storage requirements, reduces training and testing time of the prediction model [5, 8]. Finally, it helps in better understanding the processes that generate the data. However, these benefits are achieved at the cost of the additional complexity in the modeling process. With feature selection, the classification model is optimized with reduced optimal feature subset instead of the full feature set. The feature selection techniques can be broadly divided into three categories called filter, wrapper and embedded techniques [8]. The filter based feature selection techniques search for the most promising feature subset of the original set of the features based on certain evaluation function. The filter techniques work independent of the learning algorithm. Filter techniques are generally preferred to wrapper techniques because of their usability with alternative classifiers, computational efficiency and simplicity [8, 25, 16, 7]. In the literature, the important filter techniques proposed are Relief technique [29] and CFS technique [24]. The wrapper techniques select the features by using the prediction performance of the learning algorithm [12]. The wrapper techniques are less generative and computationally expensive for high dimensional data [12, 8]. The embedded techniques involve the integration of filter feature selection techniques with the learning process for a given learning algorithm [8]. Embedded techniques perform feature selection in the process of training and are usually specific to a given learning algorithm. The example of the embedded technique is C4.5 [11].Formally, the problem of feature selection can be defined as to find a subset of M features that optimally characterize the class variable T. For a given input data D tabled as N number of instances and M number of features

$X=x_i, i=1,2,...,M$ and target class variable T. The optimal characterization of the class variable depends upon two factors. The first factor is the searching algorithms that search the best subset of features meeting the optimal characterization condition. In spite of exhaustive searching, many other techniques like forward search, backward elimination, sequential forward floating search, etc. have been proposed in the literature [10, 9]. Second factor is the condition that defines the optimal characterization. Generally, the condition is minimal classification error rate or maximal of dependency of the class variable on the subset of features.

Article overview: following this introduction, section 2 highlights the important studies of mutual information (MI) based feature selection for intrusion detection. The section also presents the basic concepts and related work in the field. Finally, the paper concludes the current scenario of MI based feature selection techniques in general and especially for intrusion detection.

2 MUTUAL INFORMATION (MI) BASED FEATURE SELECTION

Feature selection:Mutual information based In the literature, many techniques have been proposed for a filter based feature selection. For any filter based feature selection technique, there are four essential steps:

1. Subset generation
2. Evaluation
3. Stopping criterion
4. Validation

Subset generation involves the searching process that generates candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one according to evaluation criteria. If the new subset produces better results, it replaces previous best one. This process is repeated until a given stopping condition is satisfied [6]. These techniques differ in the evaluation process of relevance of the features. The evaluation process may consist of subset selection or feature ranking. Subset selection process selects the relevant features and discards the irrelevant feature. Whereas, feature ranking process ranks the feature in a certain order of degree of relevance [8, 20]. The feature ranking process determines the importance of the individual features, and it neglects possible interaction of the features. In literature, many metrics have been suggested to measure the relevance of features. [6] divided the evaluation measures into five classes: 1) distance, 2) information (or uncertainty), 3) dependence, 4) consistency, and 5) classifier error rate.

1. Distance measures: It is also known as separability, divergence, or discrimination measure. For a two class problem, a feature f_i is preferred to another feature f_j. If f_i induces a greater difference between the two-class conditional probabilities than f_j; if the difference is zero then f_i and f_j are indistinguishable. Distance measure is employed in [11, 13].

2. Information measures: These measures typically determine the information gain from a feature. The information gain from a feature f_i is defined as the difference between the prior uncertainty and expected posterior uncertainty using f_i. Feature f_i is preferred to feature f_j if the information gain from feature f_i is greater than that of feature f_j. An example of this type is entropy. Information measure is employed by [3] and [26].

3. Dependence measures: Dependence measures or correlation measures quantify the ability to predict the value of one variable from the value of another variable. The correlation coefficient is a classical dependence measure and can be used to find the correlation between a feature and the class variable. If the correlation of feature f_i with class variable C is higher than the correlation of feature f_j with C, then feature f_i is preferred to f_j. A slight variation of this is to determine the dependence of a feature on other features. This value indicates the degree of redundancy of the feature. All evaluation functions based on dependence measures can be classified as distance and information measures. But, these are still kept as a separate category because, conceptually, they represent a different viewpoint. Dependence measure is employed by [22].

4. Consistency measures: This type of evaluation measures is characteristically different from the other measures because of their heavy reliance on the training dataset and use of Min-Features bias in selecting a subset of features. Min-Features bias prefers consistent hypotheses definable over as few features as possible. These measures find out the minimal size subset that satisfies the acceptable inconsistency rate that is usually set by the user. Consistency measure is employed by [1] and [18].

The above types of evaluation measures are known as "filter" techniques because of their independence from any particular classifier that may use the selected features output of the feature selection technique.

5. In contrast to the above filter techniques, there are also classifier error rate measures (also called wrapper techniques), that is used by a classifier is used for evaluating feature subsets [12]. As the features are selected using the classifier that later uses these selected features in predicting the class labels of unseen instances, the accuracy level is very high although computational cost is rather high compared to the other measures.

These metrics are used to measure the correlation between two variables. There are two types of correlation, i.e.

linear and nonlinear. The linear correlation can be measured by using a linear correlation coefficient, least square regression error and maximal information compression index. But, the linear correlation is not assumed among the features of real world data [29]. The nonlinear correlation can be measured using many different metrics. Many researchers used entropy as information theory based metric to measure the nonlinear correlation between the features. Because, entropy is a better metric to measure the uncertainty of the feature [19]. Entropy measures the uncertainty between two random variables. So, entropy based Mutual Information (MI) metric can be used to represent the dependencies of features effectively [19, 23]. MI is one of the information metric used to measure the relevance of features taking into account the higher order statistical structures existing in the data. Many researchers proposed feature selection techniques based upon MI on different evaluation functions for measuring the relevance of features [12, 16, 17, 10, 19, 23].

2.1 Basic concepts

Here, in this section preliminaries of mutual information are described. Basic definitions are given for better understanding of various feature evaluation functions in terms of mutual information.

Information theory was initially developed to find fundamental limits on compressing and reliably communicating the data [19]. Here, entropy is used as a key measure of information. It is capable to measure the uncertainty of random variables quantitatively and the amount of information shared by them effectively.

Let X be a random variable with discrete values, its entropy H (X) can be computed as

$$H(X)=-\sum_{x\in X} p(x)log(p(x)) \qquad (1)$$

where p(x) is the probability density function for X.

Joint entropy of X and Y variables is defined as

$$H_j(X,Y)=-\sum_{y\in Y}\sum_{x\in X} p(x,y)log(p(x,y)) \qquad (2)$$

where p(x, y) gives the joint probability of X and Y random variables.

The entropy of X variable after observing the values of another variable Y is called conditional entropy and is defined as

$$H_c(X|Y)=-\sum_{y\in Y}\sum_{x\in X} p(x,y)log(p(x|y)) \qquad (3)$$

where $p(x|y)$ is the posterior probabilities of X variable given the values of Y variable and p (x, y) give the joint probability of X and Y random variables.

The mutual information is defined as the amount of information shared by two variables. For variables X and Y, it is computed as

$$MI(X,Y) = -\sum_{y\in Y}\sum_{x\in X} p(x,y)log(\frac{p(x,y)}{p(x)p(y)}) \qquad (4)$$

where p(x, y) gives the joint probability of X and Y random variables and p(x), p(y) are the probability density functions of variable X and Y respectively. A large value of MI signifies high correlation of two variables. Zero value indicates that two variables are not correlated.

Conditional MI is defined as the amount of information shared by two variables when the third is known. The conditional MI between variables X and Y given Z is computed as

$$MI(X,Y|Z)=H_c(X|Z)-H_c(X|Y,Z) \qquad (5)$$

This gives the information added by Y about X which is not contained in Z.

2.2 Related work

Feature selection:Mutual information based:Related work In the literature, many researchers used MI to measure the correlation or relevance of two or more variables [2, 28, 16, 27, 7, 23, 19, 21]. A review of representative studies of MI based feature selection techniques is presented in chronological order as below:

[2] proposed a technique called MIFS (Mutual Information based Feature Selection) that utilized MI to reduce the number of features. He suggested that a good set of features are not relevant individually but also non redundant with respect to each other. That means features should be highly correlated with target class variable and not be correlated with each other. The evaluation function used for selection of feature subset was

$$EvalFunc=MI(X_n,Y)-\beta^* \sum_{k=1}^{n} MI(X_n,Y_k) \qquad (6)$$

The first factor of expression gives the feature relevance and second factor measures the penalty for correlation of the feature with each other. Here, MI ( ) is a function to compute the mutual information between two random variables. The β is the parameter to be determined empirically that varies between 0 and 1. Putting β equal to zero assumes that features are independent. Greater values of β put emphasis on reducing inter feature correlation. The authors used the tradeoff between relevance and redundancy but ignored class conditional interaction term between the features. Another issue, that is noticeable here that it evaluates features from the view of the individual, not the whole. The assigning appropriate value of β is also a critical task.

[28] proposed an evaluation function based upon the joint mutual information. The evaluation function was:

$$EvalFunc= \sum_{k=1}^{n} MI(X_nX_k,Y) \qquad (7)$$

The authors used the information between class variable and joint random variable. The joint random variable was obtained by pairing the candidate feature with already selected features. The author ignored the class conditional interaction information term again.

[16] proposed a technique called MIFS-U which is an improvement over the technique proposed by [2]. MIFS-U technique suits the systems where information is uniformly distributed. The evaluation function used was:

$$EvalFunc = MI(X_n,Y) - \beta * \sum_{k=1}^{n} \frac{MI(X_k,Y)}{H(X_k)} * MI(X_n,X_k) \qquad (8)$$

[27] proposed an evaluation function that measures the gain of combining the candidate feature with already selected features. The feature having a minimum gain with the already selected feature was added to the final subset of features. The evaluation function used was:

$$EvalFunc=MIN_k(MI[(X_n*X_k,Y)-MI(X_k,Y))] \qquad (9)$$

[7] proposed an evaluation function based on the conditional mutual information maximization. The proposed evaluation function measures the information between the candidate feature and the class conditioned on already selected features. The evaluation function expression used was:

$$EvalFunc=MIN_k[(MI(X_n,Y|X_k)] \qquad (10)$$

[23] proposed an evaluation function based upon the concept of maximum relevance minimum redundancy (MRMR). The evaluation function used was:

$$EvalFunc=MI(X_n,Y)-\frac{1}{n-1}* \sum_{k=1}^{n} MI(X_n,X_k) \qquad (11)$$

The above expression is similar to the Equation 6 proposed by [2] in MIFS, where $\beta=\frac{1}{n-1}$.

This technique tightly binds up with a specific classifier to improve performance of the classifier.

[19] proposed a dynamic mutual information based feature selection technique using a special decision tree. The proposed technique used MI as the metric for feature evaluation. They observed that the mutual information is estimated on the whole sampling space in traditional feature selection techniques. This, however, can- not exactly represent the relevance among features. They proposed a technique that re-computes the MI from unclassified instances of an unselected set of features for each addition of a feature to a subset of the selected features. The proposed technique minimizes the redundancy of the features by computing MI dynamically. But, this technique ignores class conditional interaction information term for the interaction of the candidate feature with already selected features.

[21] proposed a filter feature selection technique based on feature clustering. This technique built a dissimilarity space using information theoretic measures, in particular conditional mutual information between the features with respect to a relevant variable that represents the class labels. Hierarchical clustering was performed by computing distance based on MI between instances and centroid of the clusters. They proposed maximal-relevant-minimal-redundancy criterion based function. The proposed technique outperformed the different state of art feature selection techniques in the task of classification from the point of view of classification accuracy. The technique efficiently implemented the concept of maximal relevance and minimal redundancy to minimize classification error. But, it ignored the feature interaction.

[15] proposed a dynamic mutual information based feature selection technique using a special decision tree as an extension to study proposed in [19]. The proposed technique used MI as the metric for feature evaluation. They observed that the mutual information is estimated on the whole sampling space in traditional feature selection techniques. This, however, can- not exactly represent the relevance among features. They proposed a technique that re-computes the MI from unclassified instances of an unselected set of features for each addition of a feature to a subset of the selected features. The proposed technique minimizes the redundancy of the features by computing MI dynamically and also consider class interaction information among the features.

It can be concluded from the above cited discussion that an efficient feature selection technique should select the most relevant features by taking into account the relevance, redundancy and interaction information of the features with respect to the class. The findings of the literature review cited above can be summarized as below and are depicted in Table 1

1. Most of the researchers used the whole sample space to compute MI and use their same values throughout the whole process of feature selection [14, 19]. Here, they assume that values of information metrics remain unchanged throughout the whole computations. But, it is not true. As data instances predicted by the current feature subset cannot contribute to compute the relevance of features for unpredicted instances. So, in order to compute the relevance of the current set of features, value of MI should be recomputed only from unpredicted instances.

2. Some researchers proposed to consider the relevance of features during the feature selection process. But, they

ignore redundancy of features. For example, if two features are highly relevant but correlated. Selection of both the features will result in the selection of the redundant features. Processing of the redundant features increases the computational overhead.

3. Some researchers considered relevance and redundancy but ignored the class conditional interaction information. Class conditional interaction information is the information added by the candidate feature to already selected feature set under the condition of the given class. One feature may be irrelevant but become relevant when considered in the presence of other features. For example, one variable is irrelevant to predict the output of the XOR function of two variables, but it becomes relevant when predicted in the presence of another variable.

The table 1 summaries the findings of above sections.

Table 1: Comparative summary of MI based feature selection techniques

Study	Relevance	Redundancy	Interaction information	Dynamic computation for MI
[2]	√	√	×	×
[28]	√	√	×	×
[16]	√	√	×	×
[27]	√	√	×	×
[7]	√	√	×	×
[23]	√	√	×	×
[19]	√	√	×	√
[21]	√	√	×	NA
[15]	√	√	√	√

3 CONCLUSIONS

The aim of this paper is to present an updated review of mutual information based feature selection techniques. This text introduced need and significance of feature selection. It highlighted the use of the mutual information in filter based feature selection and different functions for evaluation of features for selecting the features. The findings of the paper are that an efficient feature selection technique selection the features based on dynamic mutual information by considering relevance, redundancy and class conditional interaction information about the features.

REFERENCES

[1] Almuallim, H., Dietterich, T.: Learning boolean concepts in the presence of many irrelevant features. Artificial Intelligence **69**(1), 279–305 (1994)

[2] Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks **5**(4), 537–550 (1994)

[3] Bell, D., Wang, H.: A formalism for relevance and its application in feature subset selection. Machine learning **41**(2), 175–195 (2000)

[4] Bellman, R.: Adaptive control processes: a guided tour princeton university press. Princeton, New Jersey, USA (1961)

[5] Daelemans, W., Hoste, V., De Meulder, F., Naudts, B.: Combined optimization of feature selection and algorithm parameters in machine learning of language. Machine Learning: ECML 2003 pp. 84–95 (2003)

[6] Dash, M., Liu, H.: Consistency-based search in feature selection. Artificial intelligence **151**(1), 155–176 (2003)

[7] Fleuret, F.: Fast binary feature selection with conditional mutual information. The Journal of Machine Learning Research **5**, 1531–1555 (2004)

[8] Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. The Journal of Machine Learning Research **3**, 1157–1182 (2003)

[9] Jain, A., Duin, R., Mao, J.: Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(1), 4–37 (2000). 10.1109/34.824819

[10] Jain, A., Zongker, D.: Feature selection: Evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(2), 153–158 (1997)

[11] Kira, K., Rendell, L.: A practical approach to feature selection. In: Proc. of the ninth international workshop on Machine learning, pp. 249–256. Morgan Kaufmann Publishers Inc. (1992)

[12] Kohavi, R., John, G.: Wrappers for feature subset selection. Artificial intelligence **97**(1-2), 273–324 (1997)

[13] Kononenko, I.: Estimating attributes: Analysis and extension of relief. In: Proc. of European Conference on Machine Learning, Catania, Italy, pp. 171–182 (1994)

[14] Kumar, G., Kumar, K.: A novel evaluation function for feature selection based upon information theory. In: Proc. of 24th Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 000,395–000,399. IEEE (2011)

[15] Kumar, G., Kumar, K.: An information theoretic approach for feature selection. Security and Communication Networks **5**(2), 178–185 (2012). 10.1002/sec.303

[16] Kwak, N., Choi, C.: Input feature selection for classification problems. IEEE Transactions on Neural Networks **13**(1), 143–159 (2002)

[17] Lin, D., Tang, X.: Conditional infomax learning: an integrated framework for feature extraction and fusion. Computer Vision– ECCV pp. 68–82 (2006)

[18] Liu, H., Setiono, R.: A probabilistic approach to feature selection-a filter solution. In: Proc. of Machine learning-international workshop then conference, pp. 319–327. Citeseer (1996)

[19] Liu, H., Sun, J., Liu, L., Zhang, H.: Feature selection with dynamic mutual information. Pattern Recognition **42**(7), 1330–1339 (2009)

[20] Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering **17**(4), 491–502 (2005)

[21] Martínez Sotoca, J., Pla, F.: Supervised feature selection by clustering using conditional mutual information-based distances. Pattern Recognition **43**(6), 2068–2081 (2010)

[22] Modrzejewski, M.: Feature selection using rough sets theory. In: Proc. of Machine Learning: ECML-93, pp. 213–226. Springer (1993)

[23] Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on **27**(8), 1226–1238 (2005)

[24] Peng, T., Leckie, C., Ramamohanarao, K.: Survey of network-based defense mechanisms countering the dos and ddos problems. ACM Computing Surveys (CSUR) **39**(1), 3 (2007)

[25] Ramaswami, M., Bhaskaran, R.: A study on feature selection techniques in educational data mining. Arxiv preprint arXiv:0912.3924 (2009)

[26] Sheinvald, J., Dom, B., Niblack, W.: A modeling approach to feature selection. In: Proc. of 10th International Conference on Pattern Recognition, vol. 1, pp. 535–539. IEEE (1990)

[27] Vidal-Naquet, M., Ullman, S.: Object recognition with informative features and linear classification. pp. 281–288. ICCV (2003)

[28] Yang, H., Moody, J.: Feature selection based on joint mutual information. In: Proc. of International ICSC symposium on advances in intelligent data analysis, pp. 22–25. Citeseer (1999)

[29] Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research **5**, 1205–1224 (2004)