CHAPTER-2

FEATURE SELECTION TECHNIQUES

2.1. INTRODUCTION

Dimensionality reduction through the choice of an appropriate feature subset selection, results in multiple uses including performance upgrading, reducing the curse of dimensionality, promoting generalization abilities, speed-up by depreciating computational power, growing model strength and lesser costs by avoiding “expensive” features.

Feature selection or variable selection is taken as the process employed for selecting the finest subset from the original features, primarily to solve the high dimensionality problem and avoid redundant and irrelevant features. There are plenty of advantages in using feature selection: reduction in training time, growing generalization performance, better classification results and excellent model interpretability. Although best classifiers are used for classification process there is no assurance in getting better results. So, Feature selection when combined with good classifiers there is more probability in getting good results [10][11].

Feature selection varies from Feature extraction that generates new features by merging the original features. While Feature selection, is responsible in maintaining prototypes of the selected features, which is well-suited in many fields. It is imperative and usually used Pre-Processing
techniques, applied to the normalized training dataset and results in reduced feature set selection.

There are different feature selections algorithms which have achieved good results in many applications. They have the following objectives [12].

1. Identify the feature subset that is basic and satisfies the target concept [13].

2. Choose a subset of L features from a set a K features, L<M, in such a way that the criterion function is maximized [14].

3. Select a subset of features for increasing prediction accuracy or reducing the size of the structure without considerably decreasing prediction accuracy of the classifier constructed using only the selected features [15].

2.2. FEATURE SELECTION PROCEDURE

Generally feature selection procedure encompass four essential steps: 1) subset generation; 2) subset evaluation; 3) terminating criterion and 4) result validation [16]. The approach starts with subset creation which uses scrupulous exploration trick to generate candidate feature subsets. Furthermore every candidate subset is assessed based on a precise estimation principle and differentiated with the earlier finest ones. If it is stronger among all others, then it substitutes the best preceding one. This procedure of subset establishment and assessment is performed repetitively, and stopped when a specified terminating condition is fulfilled. Lastly the chosen top feature subset is
confirmed by earlier information. Eventually, search pattern and estimation criterion are two key issues in the research of feature selection. The General Feature Selection Procedure is shown in fig 2.1.

![Feature Selection Procedures Diagram](image)

**Fig. 2.1: Feature Selection Procedures**

2.3. SUBSET GENERATION

Subset generation is done in two steps. First step is finding the search direction and the other one is identifying search strategy. Search direction can be forward, backward and random. If the search begins with empty set and adding features sequentially it is called forward search. If the search begins with a full set and removing features consecutively it is called backward search. The search may also begin by selecting features randomly so that local optima can be avoided, it is called random search.

In the second step, there are different search strategies for discovering an optimal or suboptimal feature subset. For a data set with N features, there exist 2N candidate subsets. An exhaustive search approach considers the entire
2N feature subsets to trace most excellent ones. It has the complexity of $O(2N)$ given in terms of the data set dimensionality. There are other search strategies like sequential [17], complete [18], and exponential search [19] strategies.

### 2.4. Subset Evaluation

After subset generation is completed, there is a need for evaluating each generated subset. Evaluation criteria are broadly classified into two types with respect to their reliance on learning algorithms that are accordingly pertained on the resultant feature subset. They are either independent criteria or dependent criteria. A few examples of prevailing independent criteria are information, dependency, distance, consistency measures [20], [21], [22], [23]. Usually an independent principle is mainly applied in algorithms with the filter model. Therefore a dependent principle is applied in the wrapper model, needs a preplanned learning algorithm in feature selection. It utilizes behavior of the learning algorithm applied for the chosen subset for adjudicating the features to be picked. Then considering whether they rely on the learning algorithms or not, feature selection algorithms are basically partitioned into three groups: filter, wrapper, and hybrid methods.

#### 2.4.1. Filter Method

Filter methods work irrespective of any classifier; they filter out unimportant, unnecessary and noisy features using preprocessing steps before induction process starts. They give ranks to the features or feature subsets. They make use of intrinsic properties to detect feature subsets [24], [25]. Mainly features are examined by their importance and undesirable
functionalities with regard to target classes. It has many advantages over wrappers; they have excellent simplification characteristics since they are self-determining without any scrupulous learning method [26]. Filters generally evaluate feature subsets by their data content using correlation measures, distance measures, gain ratio, principal component analysis (PCA), information gain etc. Filter techniques chuck features upon their evaluation, using data characteristics or using several kinds of statistical analysis, irrespective of any learning method involved. The structure of filter approach is shown in fig 2.2.

![Filter approach diagram](image)

**Fig. 2.2: Filter approach**

Correlation based feature selection is a heuristic technique used for assessing advantages of a feature subset. Therefore, based on correlation idea, a feature is said to be outstanding, only when that feature is very much interrelated to the class, not to the remaining features [27]. So there is a need in finding correlation between features, which represent the important and highly effective characteristics.
Information Gain (IG) helps us to decide which feature is most helpful for classification, in regard to its entropy value. Entropy specifies the information content of a feature or how much information it is providing us. Therefore the higher the entropy value, the more the information content.

2.4.2. Wrapper Method

Wrapper Method exploits a learning algorithm for assessing features or feature subsets through their predictive accuracy. They need more computation time and slower because of the repeating process; however they give more accurate results than filter procedure. This is the disadvantage for high dimensional data but it could be overwhelmed by using a fast learning algorithm. Additionally, the subset assessment is strongly connected with a classifier used; wrappers must be therefore has the capability in repeating the process, whenever another classifier is employed for feature evaluation. The process of Wrapper approach is shown in fig 2.3.

![Diagram of Wrapper Method]

**Fig. 2.3: Wrapper Method**

Wrappers [28], [29], as contrasting to filter methods, investigate for the finest subset by using an empirical risk estimate for a certain classifier. Thus,
they are regulated to the particular relations amid the classification algorithms and the available training data. There are several Wrapper techniques mainly F-score, Random forest etc. Many Machine learning algorithms are used as Wrapper approaches namely Support Vector Machines, Genetic algorithms, decision tress and so on.

2.4.3. Hybrid Method

By combining the advantages of filter and wrapper approaches, hybrid models [30] [31] are constructed so that efficiency is improved and feature selection becomes faster. It typically begins initially with blank subset and iterated for getting most excellent subsets in the series of growing cardinality. Usually, a hybrid method makes use of independent strategy to choose most promising subsets for a specified cardinality and uses the classifier in selecting ultimate superior subset from all the most excellent subsets with different cardinalities. The Hybrid approach is shown in the fig 2.4.

![Diagram of Hybrid Method]

**Fig. 2.4: Hybrid Method**

While filters can provide an intellectual guideline for wrappers, like reduction in the search space, an excellent initial position, or a smaller search
path, which support in scaling hybrid techniques into high dimensionality issues.

2.4.4. Feature Ranking

There are several feature selection algorithms, among them feature ranking gains more interest in research because of their ease in application, simplicity and finest empirical results. These approaches accord score or rank to the features based on certain criteria and exploit the ranks or scores for the selection mechanism. In terms of computation, feature ranking is more efficient than remaining feature selection techniques, as they compute $M$ scores and sort all the scores accordingly. Then subsets of features are selected and applied to the respective classifier or predictor.

As the research grows in the feature selection perspective, the researchers introduced varying feature selection criteria. In a related study Hsu et al. [32] has given association between rank phenomena and score phenomena through setting up a technique known as rank/score graph. They have demonstrated that subjected to definite circumstances, rank combination exceeds score combination. In another study the authors [33] successfully exploited rank combination in merging with different feature selection techniques. Hence, rankings obtained for features are merged with a weighted summation gained with each of the entry scores achieved by separate feature selection methods. Hence the combined approach ameliorates well than individual feature selection methods in several applications.
2.5. STOPPING CRITERIA

Feature selection is terminated based on the following principles [34]:

1. When the search gets completed.
2. When the particular bound is reached, here bound is given as least number of attributes or maximum number of iteration.
3. When an estimated number of iterations are finished.
4. When an optimum feature subset in rapport with evaluation criterion.
5. When the change (adding up or elimination of features) of feature subsets will not construct a superior subset.

2.6. RESULT VALIDATION

A central means of result validation is obtained by measuring the outcome directly from previous information. As a result we depend on some indirect methods through examining the changes in the mining capability with the change of features. Ultimately we have to exploit the learning capability on the testing data as a worthiness of the chosen feature subsets. For an illustration, classification error rate is used as an efficiency decisive factor for a classification process. Then for a chosen feature subset, just analyze the “before-and-after“ analysis test for comparing the fault estimate concerned with learning task, performed on both complete set of features and selected subset [17].
2.7. GENERAL FEATURE SELECTION STRATEGIES

2.7.1. Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is an important means of unsupervised feature selection. It has been largely used for information retrieval for reducing the dimension of the document vector space. SVD is merely considered as a method of mathematical and statistical approaches in an extensive range of fields, including spectral analysis, Eigen value decomposition, image processing, information retrieval, computer vision and fractal analysis.

It is directly associated to the diagonal form given as \( A = Q \Lambda Q^T \) of a symmetric matrix. If this matrix is not symmetric, then we can factorize A by \( Q_1 D Q_2^T \) where \( Q_1, Q_2 \) are orthogonal matrices and D is non-negative and diagonal matrix with \( \{d_1, d_2, d_3, \ldots, d_n\} \) entries in it. Then the diagonal values of D are called singular values. The matrices are created in the following manner: The Columns of \( Q_1 \) (m×m) are the Eigen vectors of \( AA^T \) and the columns of \( Q_2 \) (n×n) are the Eigen vectors of \( A^T A \). The r singular values on the diagonal of D (m×n) are the square roots of the nonzero Eigen values of both \( A^T A \) and \( AA^T \) [35].

Previously work had done on employing Singular Value Decomposition and SVM in developing new IDSs by the authors Tao et.al [36]. The said method is a feature selection based made on orthogonal projection coefficients using SVD. Then the SVM mechanism is done on the newly attained feature.
vector sets. They utilized RBF kernel parameters with the grid-search and applied cross-validation in the projected work. Ultimately testing outcomes proved that this is the novel IDS method which is powerful and hold many desirable features as contrasting with lots of existing methods.

The study on SVD with IDS [37] inspects the applicability of Spectral Analysis technique called singular value decomposition (SVD) as a data processing step for decreasing the dimensionality of the data. This diminution focuses most outstanding features in the data by eliminating the noise. Here with data processing step, it makes data noise-free as well as lessening the dimensionality of the data, thus reducing time essential for computation. Then the mentioned approach will be implemented with other existing methods to enhance their efficiency. They have made experiments on various data sets like DARPA’98, UNM sendmail, inetd, and login-ps data sets for demonstrating, reduction of data dimensionality will not disgrace the efficiency of the IDS.

2.7.2. Recursive Feature Elimination

Recursive Feature Elimination (RFE) is the one that is related to wrapper approach. It is narrow in picking a small set of highly discriminative features. With the approach given in [38] they have made it successfully with merging both Support Vector Machine (SVM) and Recursive Feature Elimination (RFE). This aims at discarding features which are not performing well with the rest of the features. Relying on this aspect the authors have proposed a novel scheme of selection with Support Vector Machine methods relied on Recursive Feature Elimination (RFE). They have asserted practically
that genes chosen with these techniques attained enhanced classification efficiency and are purely associated with cancer. Comparisons done on baseline method have shown that this technique removes gene excessiveness efficiently and has given superior, denser gene subsets. It has achieved 98% accurateness when facing with baseline method with just 86% accurateness.

Greedy methods like Recursive Feature Elimination are noted by several researchers is less prevailing with over fitting than other search techniques. Even though examining theoretic analysis of the conventional characteristics of SVM-RFE is yet to be worked out.

This technique was used by the authors [39] in proposing upgraded SVM model for intrusion detection. With this new model they implemented Recursive Feature Elimination (RFE) for ranking features of intrusion detection data and k-Nearest Neighbor (KNN) for enhancing accuracy in learning process. Moreover, in training SVM they considered only significant features. It is recommended that this novel approach is effectual for the KDD cup 99 dataset. Against traditional SVM, it possesses lower false negative rates. And also, the time taken to identify an intrusion for this technique is lesser than the standard SVM. This method is being significantly good in terms of accuracy. Therefore it holds an advantage of having lesser running times since only small numbers of attributes are employed for classification.
2.7.3. Principal Component Analysis (PCA)

PCA is a technique good at recognizing patterns in data set, and projects them in the manner, through which exposed their resemblances and dissimilarities. Since patterns are tough to locate in records of higher dimensions, it is considered as a dominant examining tool. Anyway patterns are obtained; the data is condensed by shrinking the dimensionality with no considerable loss of information. The class labels are not necessary in this approach. Mainly intension of PCA lies in dealing with discovering a group of \( d \) orthogonal basis vectors which expands the relationship between the initial dimensions [40]. The process of PCA is shown in fig. 2.5.

![Fig. 2.5: The Steps in Principal Component Analysis](image-url)
In a study made by the authors [41], they have efficiently developed intrusion detection system based on PCA and Support Vector Machines (SVMs) for finding best feature subset and classification accuracy. Then they verified its efficiency and practicability of the Intrusion Detection System, through performing numerous experimentations on NSL-KDD dataset. From tested results shown in mentioned approach, they are able to make faster detection of intrusions and decrease the memory space and CPU time.

2.7.4. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is the technique helps in identifying outstanding features. It obtains the optimal transformation matrix in order to protect the majority of the data by distinguishing different classes. Therefore with this method, it utilizes class labels for formulating exactly the optimization technique. LDA is a conventional statistical technique mainly used for supervised dimensionality reduction and also for classification. It calculates most favorable transformation through decreasing the within-class distance and increasing the between-class distance within similar instance. Consequently it achieves highest class discrimination. The optimal transformation in this approach is computed by using Eigen decomposition on the scatter matrices. LDA is widely employed in many applications containing higher dimensional data. The main goals of LDA are: performing dimensionality reduction “whereas maintaining as much of the class discriminatory information as possible”, and also finds to notice directions
followed by classes that are separated efficiently. It captures both scatter within-classes and also scatters between-classes.

LDA locates the vectors in the indispensable space which differentiate well amongst classes. Then with all samples containing all classes the between-class scatters matrix $S_B$ and within-class Scatter matrix $S_W$ are defined by:

$$S_B = \sum_{i=1}^{c} (\mu_i - \mu)(\mu_i - \mu)^T$$  \hspace{1cm} (2.7.4.1)

$$S_W = \sum_{i=1}^{c} \sum_{j=1}^{M_i} (Y_{ij} - \mu_i)(Y_{ij} - \mu_i)^T$$  \hspace{1cm} (2.7.4.2)

Where $M_i$ is the number of training samples in class $i$, $C$ is the number of distinct classes, $\mu_i$ is the mean vector of samples belonging to class $i$ and $Y_{ij}$ represents the set of samples belonging to class $i$ with $Y_j$ being the jth data of that class. $S_w$ represents the scatter of features around the mean of each class and $S_B$ represents the scatter of features around the overall mean for all classes. The goal is to maximize $S_B$ while minimizing $S_W$, in other words, maximize the ratio

$$\frac{\text{det} | S_B |}{\text{det} | S_W |}$$  \hspace{1cm} (2.7.4.3)

The ratio is maximized when the column vectors of the projection matrix are the Eigen vectors of $S_w^{-1} S_B$. In order to prevent $S_w$ to become singular, Information Gain is used as preprocessing step. The sequence of steps carried out in LDA is shown in fig 2.6.
Fig. 2.6: The Process of Linear Discriminant Analysis

It has been effectively exploited as a good feature selection technique and considerably achieved good results. Experimental results shown that this approach gives superior and strong representation of data, since it was capable in decreasing features with 97% data reduction and around 94% time reduction in training with approximately same accuracy achieved in detecting new attacks [42].

-48-