NJESR/ April-2023/Volume-5/Issue-4 E-ISSN-2582-5836
DOI-10.53571/NJESR.2023.5.4.1-13

Machine Learning Feature Extraction Application

Sreevidya V.S
Lecturer In Computer Engineering,
Government Polytechnic College
Punalur

(Received:1March2023/Revised:19March2023/Accepted:29March2023/Published:4April2023)

Abstract

Dimensionality reduction is a useful machine learning preprocessing step that removes irrelevant and redundant data, improves learning accuracy, and makes results easier to understand. However, the recent rise in data dimensionality poses a significant obstacle to the efficiency and effectiveness of numerous existing methods for feature selection and extraction. Dimensionality reduction is a significant area in machine learning and pattern recognition, where numerous approaches have been proposed. This paper examines some well-known methods for selecting features and extracting features to see how well they can be used to boost the performance of learning algorithms and, consequently, the classifier's predictive accuracy. It is presented an effort to briefly analyze dimensionality reduction techniques with the intention of examining the advantages and disadvantages of some widely used dimensionality reduction techniques.

Keywords: Age Related Macula Degeneration (AMD); Selection Of Features; Selecting A Subset of Features; Transformation And Extraction of Features; FSA's; Correlation-Based Method, RELIEF; PCA; ICA.

Introduction

In high-dimensional data analysis, visualization, and modeling, dimension reduction is widespread preprocessing. Feature Selection is one of the simplest methods for reducing dimensionality; one chooses just those info aspects that contain the applicable data for taking care of the specific issue. A more general approach called "feature extraction" aims to transform the input space into a low-dimensional subspace that keeps the majority of the relevant information[1]. With the intention of enhancing performance, such as estimated accuracy, visualization, and comprehension of learned knowledge[2], feature extraction and selection methods are utilized singly or in combination. The following are typical subcategories of features: relevant, redundant, or irrelevant. During the process of feature selection, a subset of the data's available features is chosen for the learning algorithm process. The subset with the

fewest dimensions that most influence learning accuracy is the best[3]. The advantage of feature selection is that important information about a single feature is not lost. However, if only a small number of features are needed and the original features are very different, some information may be lost because some features must be left out. On the other hand, with dimensionality reduction, which is also known as feature extraction, the size of the feature space can frequently be reduced without compromising the information that was present in the original feature space. The fact that the linear combination of the original features is typically unintelligible and that information about how much an original feature contributes is frequently lost is one disadvantage of feature extraction[4]. mRmR, RELIEF, CMIM, Correlation Coefficient, BW-ratio, INTERACT, GA, SVM-REF, PCA (Principal Component Analysis), Non-Linear Principal Component Analysis, Independent Component Analysis, and correlation based feature selection are all approaches that merit mentioning. Studies show that a lot of work has been put into developing the best methods for feature selection and extraction. Due to the large number of existing algorithms for feature selection and extraction, it is necessary to rely on criteria that allow for appropriate algorithm selection in specific circumstances. Based on experiments carried out by researchers to examine how these techniques contribute to improving the predictive accuracy of classification algorithms, a brief review of these methods is carried out to determine the suitability of various feature selection and feature extraction techniques in particular circumstances. In this study, we introduce readers to various methods for dimensionality reduction.

Literature Review

Techniques for reducing dimension have emerged as a pressing need in the medical field (automated application). Today, an immense measure of information is created in the clinical space. This includes a patient's possible symptoms as well as the results of numerous medical tests. Attributes and input variables are synonymous with feature. Examples of features in medical diagnosis include symptoms, which are based on a set of variables that classify a patient's health status (such as the symptoms of dry or wet age-related macular degeneration (AMD) in diabetic retinopathy). A review of some widely used feature selection and feature extraction techniques for ophthalmologists to use in the detection and diagnosis of a variety of eye diseases, including glaucoma, diabetic retinopathy, and especially the automatic detection of age-related macular degeneration, is provided in this section. This review's primary objective is to educate practitioners about the advantages of and, in some instances, the necessity of using

dimensionality reduction techniques. To get benefit from dimensionality decrease strategies to boost precision of learning calculation, there is need to have familiarity with different benefits of these methods. L. Ladha et al in[3] have been offered following benefits of component determination:

• It lessens the dimensionality of the element space, to restrict capacity prerequisites and speed up.

• It eliminates noisy, redundant, or irrelevant data.

• Speeding up the learning algorithms' execution time is one of the immediate effects for tasks involving data analysis.

• Increasing the accuracy of the resulting model and improving the quality of the data

• Reducing the feature set to save resources for the subsequent data collection or use

• Enhancing performance to increase predictive accuracy

• Data understanding to learn about the process that created the data or to simply understand how the data is displayed.

An approach for extracting image-based features from digital retinal images to classify AMD has been proposed by P. Soliz and colleagues[6]. An ophthalmologist has divided 100 images into 12 categories based on the disease's visual characteristics. In order to extract features and provide input for a classifier, independent components analysis (ICA) was utilized. It has been demonstrated that ICA can implicitly extract the mathematical features from each image to define the phenotype and robustly detect and characterize features in funds images. The effects of class noise—also known as misclassification or mislabeling—on supervised learning in medical domains have been examined by M. Pechenizky[7]. It has been suggested to use feature extraction as a pre-processing step to lessen the impact of class noise on the learning process after a review of related work on learning from noise data. The filtering methods specifically deal with noise. Researchers have acknowledged the usefulness of numerous filtering strategies that have been compiled in this summary. However, the same researchers have discovered some real-world issues with filtering strategies. One concern is that without the assistance of an expert, it is difficult to distinguish noise from exceptions (outliers). Another worry is that a filtering method can use an expected level of noise as an input parameter, but this value is rarely known for a given dataset. Because it aids in implicitly preventing overfitting within learning techniques, feature extraction methods employing PCA work better with noise tolerance

techniques. The negative impact of mislabeled instances in the data can be reduced by using feature extraction techniques prior to supervised learning. A singular value decomposition (SVD) and principle component analysis (PCA)-based diabetic detection method with ANN and a feature set has been proposed in[8]. The findings of the experiments indicate that the composition of ANN-SVD+PCA is a trustworthy and accurate method for detecting diabetes that requires less computational power. Because of the noisy data, feature extraction methods were found to be much better suited for the automated detection of diseases by ophthalmologists than feature selection methods. because the majority of biomedical datasets contain noisy data rather than irrelevant or redundant data.

Dimensionality Reduction Approaches

High layered information is risky for characterization calculations because of high computational expense and memory utilization[4]. The two methods for reducing dimensionality are feature selection (FS) and feature extraction (also known as dimensionality reduction explicitly or feature transformation). The benefit of FS is that there is no loss of information about the importance of a single feature; however, if only a small number of features are required and the original features are extremely diverse, information may be lost because some of the features must be omitted during the process of selecting the feature subset. However, in feature extraction, the size of the feature space can frequently be reduced without sacrificing a significant amount of the original feature space's information. Depending on the type of data and application domain, feature extraction or selection methods should be chosen.

Feature Selection

High-dimensional data contains features that may be irrelevant, misleading, or redundant. This increases the size of the search space, makes it harder to process the data further, and does not help the learning process. The process of selecting the best features from all those that can be used to differentiate classes is known as feature subset selection. A computational model called the feature selection algorithm (FSA) is sparked by a particular definition of relevance. An empirical comparison of different feature selection algorithms has been provided by L. Ladha et al.[3]. According to some evaluation criteria, feature selection is generally referred to as a search problem. The following are some characteristics of feature selection algorithms: three sorts of search is conceivable remarkable, successive, or irregular. (ii) A subset of the generation of successors: five distinct administrators can be considered to create replacement that are;

Compound, Weighted, Random, Forward, and Backward iii) Criteria for Evaluation: As depicted in Figure 2, the Probability of Error, Divergence, Dependence, Interclass Distance, Information, or Uncertainty and Consistency Evaluation can be used to measure evaluation of successors. There are three main types of feature selection methods: wrappers, filters, and an embedded or hybrid method Because the feature selection process is optimized for the classifier that will be used, wrappers methods perform better than filter methods. Notwithstanding, covering techniques have costly to be utilized for enormous element space due to high computational expense and each list of capabilities should be assessed with the prepared classifier that at last make include determination process slow. Compared to wrapper methods, which are better suited for high-dimensional data sets and have a lower computational cost, filter methods are faster but have poor classification reliability. Recently developed hybrid/embedded methods combine the advantages of wrappers and filters. An independent test and a performance evaluation function of the feature subset are used in a hybrid approach[10]. Channels techniques can be additionally sorted into two gatherings, to be specific component weighting calculations and subset search calculations as displayed in Figure 1. Individual features are given weights by feature weighting algorithms, which then rank them according to how closely they relate to the intended concept [11].

Feature Extraction/Transformation

The process of extracting features involves changing some of the original features in order to produce other features that are more significant. The following is how Brian Ripley defined feature extraction: The construction of linear combinations "Tx of continuous features with good discriminatory power between classes" is typically referred to as feature extraction. Finding a suitable representation for multivariate data is a significant challenge in both research on neural networks and other fields like artificial intelligence. In this setting, features extraction can be used to simplify data representation by representing each variable in feature space as a linear combination of the initial input variable. Principle Component Analysis (PCA), which Karl developed, is the feature extraction method that is used the most and is most frequently. PCA has been the subject of numerous proposals. PCA is a straightforward nonparametric method for extracting the most relevant data from noisy or redundant data. PCA is a linear transformation of data that maximizes information while reducing redundancy (measured by covariance) and redundancy (measured by variance). Two distinct types of data sets—e-mail data and drug

discovery data—were used to empirically test the relationship between various dimensionality reduction methods, including feature subset selection with information gain (IG) and wrapper methods, feature extraction with deferent flavors of PCA methods, and their effects on classification performance. According to the findings, PCA feature extraction (transformation) is highly dependent on the kind of data. For both types of data, the feature selection method Wrapper has a smaller impact on classification accuracy than IG. The significance of a dimensionality reduction procedure is emphasized by the experimental results. When compared to feature extraction methods, wrappers methods for feature selection typically produce the smallest feature subsets with very comparable classification accuracy. However, feature extraction methods have lower computational costs than wrappers. In order to improve classification performance, Veerabhadrappa and L. Rangarajan proposed bi-level dimensionality reduction methods that integrate feature selection and extraction techniques. In the first level of dimensionality reduction, they have proposed two methods: Mutual correlation is used to select features. PCA and LPP are used to extract features from selected features at the second level. The proposed method was tested on a number of standard datasets to see how well it worked. The obtained results demonstrate that the proposed system performs better than single-level dimensionality reduction methods.

Analysis Of Different Feature Selection And Feature Extraction Techniques

Feature Selection Algorithms (FSA)

L. Ladha et al.[3] provided an overview of some fundamental FSAs and their limitations. The most widely used statistical test for determining deviation from the expected distribution when assuming that the feature occurrence is actually independent of class values is the chi-squared test. The root-of-square differences between two objects' coordinates are the subject of the Euclidian Distance investigation. This method has the advantage of not affecting the distance when new objects, which might be outliers, are added to the analysis. Notwithstanding, Euclidian distance can be enormously impacted by contrasts in scale among the aspect from which the distance is figured. The t-test determines whether two groups' averages are statistically distinct from one another. When comparing the average of two groups, this analysis is highly recommended, and it is especially suitable for the posttest-only two-group randomized experimental design. When a feature is present versus absent, the increase in entropy is measured by Information Gain (IG). Measurement of informational entropy, a more general method, is

used here to solve the problem of deciding how important a feature is in feature space. CorrelationBased Feature Selection (CFS) looks for features in a subset based on their degree of redundancy. The goal of the evaluation process is to find subsets of features that are highly correlated with the class on their own but have a low correlation between them. Importance of gathering of elements develops with the connection among's highlights and class, and diminishes with developing between relationship. CFS is typically used in conjunction with forward selection, backward elimination, bi-directional search, best first search, and genetic search to identify the best feature subset. In Lei Yo et al., a novel concept called predominant correlation was introduced, as well as a fast filtering method that could identify relevant features and redundancy among relevant features without the use of pairwise correlation analysis. The simplest greedy search algorithm is Sequential Forward Selection, or SFS. When the optimal subset contains a small number of features, SFS performs best. The inability of SFS to remove features that become obsolete as other features are added is its primary drawback. In contrast to SFS, Sequential Backward Elimination (SBE) operates in reverse. When there are a lot of features in the feature subset, SBE work best. The inability of SBE to reevaluate a feature's usefulness after it has been discarded is its primary limitation. A variant of SFS and SBE, Plus-L Minus-R Selection (LRS) is a generalization. It makes an effort to make up for the shortcomings of SFS and SBE by offering some backtracking capabilities. The main problem with individual feature selection algorithms is that they only capture the relevance of features to the target concept and avoid redundancy among features. This is the problem with individual feature selection algorithms. According to empirical evidence from the literature on features selection, redundant features, in addition to being irrelevant, also have an impact on the speed and accuracy of learning algorithms and ought to be eliminated. Therefore, pure relevance-based feature weighting algorithms do not adequately satisfy the requirement for feature selection in the context of feature selection for high-dimensional data, where there may be numerous redundant features[4].

Characteristics Of Feature Subset Selection

Subset search algorithms look for candidate feature subsets using a particular evaluation measure to determine how good each subset is. Using a variety of public domain datasets, the performance of various feature selection algorithms has been examined. With a few widely used methods, the number of reduced features and their impact on learning performance have been

measured, compared, and evaluated. To describe target conceptions of the learning process, a feature selection method should select the best feature subset from feature space. When selecting features, the following factors must be taken into account: 1. Beginning Point, 3. Search Methodology 4. Evaluation of a Subset Stopping Points Table 1 presents a comparative analysis of feature selection methods based on these aspects. We defined feature selection techniques in such a way that we could provide an overview of comparative analysis regarding the search organization, feature generation, and evaluation measure that each feature selection technique implies. This overview can help practitioners in the field choose a technique that is appropriate for their goals and resources. Nine different methods for selecting features have been discussed. The mutual information (MI) of two random variables is utilized in mRMR, or Minimal Redundancy and Maximal Relevance. A measure of the mutual dependence of the two variables is called MI.

MI between features is used as the redundancy of each feature, and MI between classes is used as the relevance of the feature to the class. The well-known weighting (ranking) method known as I-RELIEF measures the relevance of features in neighboring samples that are located around target samples. Relief measures the distance between the target and hit samples after locating the closest sample in the same category's feature space, known as the hit sample. Additionally, it performs the same work by locating the nearest sample, or miss sample, in the opposite category. The weight of the target feature is the difference between those measured distances. Various variations of this fundamental algorithm were extracted. The I-RELIEF approach lessens the original RELIEF method's bias. Using conditional mutual information, Conditional Mutual Information Maximization (CMIM) selects a feature subset that is most relevant to the target class. Binary values are required for both feature values and output classes in CMIM. The Correlation Coefficient approach looks at how well each feature helps to separate classes. Using the mean and standard deviation of each feature across all samples from both classes, ranking criteria are used to rank the features. The Between-Within Ratio (BW-ratio) method selects the feature with the highest BW-ratio by comparing sums of squares from within and between groups for that feature. Based on its correlation to the class, a single variable may be deemed irrelevant; however, when combined with other features in the feature space, it becomes extremely relevant[7]. Features interact in the INTERACT methods. Backward elimination and the measurement of consistency contribution are used in this method to identify interacting

features. The C-consistency of a feature is a measure of how much a feature's removal will affect consistency; for instance, the C-consistency of an irrelevant feature will be zero. Genetic Algorithm employs a random strategy. GA are a subset of evolutionary algorithms that employ evolutionary biology-inspired methods like inheritance, mutation, selection, and crossover. In feature selection problems[8], a binary string is used to represent each feature set. Backward elimination is carried out by the wrapper method known as Recursive Feature Elimination (SVM-RFE). SVM-REF uses the weight vector w as a ranking criterion to identify the m features with the largest margin of class separation[9]. Using gene expression data and a shrunken centroid, the statistical method known as Prediction Analysis of Microarray (PAM) can predict classes. Subsets of genes that best represent the class are identified using the method of nearest shrunken centroid[10].[6] Provides an experimental evaluation of the aforementioned feature subset selection methods. Seven UCI machine learning data sets, including lung cancer, leukemia, and five others, have been taken and preprocessed for discrete feature selection.

Feature Extraction/Transformation Methods

It is crucial for the subsequent data analysis; Pattern recognition, de-noising, data compression, visualization, or any other method by which the data is presented in a way that makes it easier to analyze it are all examples. In order to locate a suitable transformation, a number of fundamental methods have been developed. In Independent Component Analysis (ICA), a linear transformation technique, the desired representation is one that minimizes the components' statistical dependence. Results from neurosciences suggest that the same principle of redundancy reduction explains some aspects of the brain's early processing of sensory data, which is why ICA is used for feature extraction. Similar to the closely related projection pursuit method, ICA can also be used in exploratory data analysis. The theory of redundancy reduction[2] serves as the impetus for the utilization of feature extraction. ICA algorithms fall into two broad categories. The reduction of mutual information is the foundation of some algorithms; The maximization of non-gaussianity is where others get their start. We are looking for maximally independent components by employing an algorithm that aims to minimize mutual information, which can be thought of as the reduction in uncertainty regarding variable X following the observation of Y. By focusing on non-gaussianity, you can estimate the independent component in a different way. Forcing each component to be as far from the normal distribution as possible is one method for extracting the components. ICA typically requires five conditions to be met: First and foremost,

the source signals must be statistically distinct; 2-The number of mixed observed signals and the number of source signals must be equal, and the mixtures must be linearly independent of one another; 3-the model should be without commotion; 4- The data needs to be centered; 5-With the exception of one signal source, which may have a Gaussian probability density function (pdf), the source signals must not have one. An orthogonal transformation is used in Principle Component Analysis (PCA) to transform samples of correlated variables into samples of linearly uncorrelated features. Principle components are the new features, and they are equal to or less than the initial variables. Since PCA is an unsupervised method, the data's label information is not included. The principle components are independent if the data are normally distributed. The principal justification for the utilization of PCA concerns the way that PCA is a straightforward nonparametric strategy used to remove the most important data from a bunch of excess or boisterous information.

By eliminating the final principle components that do not significantly contribute to the observed variability, PCA reduces the number of original variables. PCA is a linear transformation of data that maximizes information while minimizing redundancy (measured through covariance). New variables known as principle components (PCs) have two characteristics: 1) The initial variables are combined linearly in each PC; 2) There is no correlation between the PCs, and redundant data is also removed[2]. Data compression, image analysis, visualization, pattern recognition, regression, and time series prediction are the primary areas of application for PCA. There are some limitations with PCA;

1. It is predicated on the linear relationship between variables.
2. If all of the variables are assumed to be scaled numerically, then its interpretation is only reasonable.
3. It misses the mark on probabilistic model construction which is significant in numerous settings, for example, combination demonstrating and Bayesian choice.

Guttman (1941) proposed Nonlinear Principle Components Analysis as an alternative strategy for overcoming the first and second limitations. While this approach is suitable for variables with mixed measurement levels (nominal, ordinal, and numeric), it shares the same goal as PCA. Nonlinear PCA, also known as categorical PCA[23], considers all variables to be categorical, with each distinct value of a variable being referred to as a category. Standard PCA cannot analyze nominal variables. Categorical variables are another name for ordinal variables. Such variables

are made up of sorted groups, like the values on a Likert-type rating scale. The primary distinction lies in the fact that, in nonlinear PCA, the measured variables are quantified during analysis, whereas in linear PCA, the measures variables are directly analyzed. By allowing the noise component to have an isotropic structure, Probabilistic Principle Component Analysis (PPCA) overcomes the third limitation; The maximum likelihood estimation technique is used in this model's parameter learning stage, where the PCA is implicitly incorporated. Additionally, an effective expectation/maximization (EM) algorithm is developed for iterative parameter learning. Using a kernel trick, Kernel Principle Component Analysis (KPCA) overcomes the initial limitation. The main idea behind KPCA is to avoid using the kernel function to directly evaluate the required dot product in a high-dimensional feature space. As a result, no explicit nonlinear function is required to transfer the data to the feature space from the original space. Probabilistic kernel principle component analysis (PKPCA), which naturally combines PPCA and KPCA to overcome PCA's limitations, has been proposed in [2-4] as a probabilistic method for analyzing the kernel principle component.

Conclusion

A study of component determination and extraction is proposed. The reduction of feature space for better data analysis is the goal of both approaches. When real-world datasets, which may contain hundreds or thousands of features, are taken into consideration, this aspect takes on a greater significance. The principal distinction between include choice and extraction is that the first plays out the decrease by choosing a subset of highlights without changing them, while include extraction lessens dimensionality by processing a change of the first elements to make different highlights that ought to be more critical. Traditional methods, their most recent improvements, and some interesting applications for feature selection. Because it highlights the characteristics that have the greatest impact on the phenomenon under consideration, feature selection enhances knowledge of the process under consideration. In addition, the adopted learning machine's computation time and accuracy must be taken into account because they are essential for machine and data mining applications.

References

[1]. Liang, H., Sun, X., Sun, Y. and Gao, Y., 2017. Text feature extraction based on deep learning: a review. EURASIP journal on wireless communications and networking, 2017(1), pp.1-12.

[2]. Hahn, U. and Oleynik, M., 2020. Medical information extraction in the age of deep learning. Yearbook of medical informatics, 29(01), pp.208-220.

[3]. Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S. and Liu, H., 2018. Clinical information extraction applications: a literature review. Journal of biomedical informatics, 77, pp.34-49.

[4]. Freitag, D., 2000. Machine learning for information extraction in informal domains. Machine learning, 39(2-3), pp.169-202.

[5]. Dara, S. and Tumma, P., 2018, March. Feature extraction by using deep learning: A survey. In 2018 Second international conference on electronics, communication and aerospace technology (ICECA) (pp. 1795-1801). IEEE.

[6]. Khalid, S., Khalil, T. and Nasreen, S., 2014, August. A survey of feature selection and feature extraction techniques in machine learning. In 2014 science and information conference (pp. 372-378). IEEE.

[7]. Bonet-Solà, D. and Alsina-Pagès, R.M., 2021. A comparative survey of feature extraction and machine learning methods in diverse acoustic environments. Sensors, 21(4), p.1274.

[8]. Xu, X., Zhou, Y. and Leng, Y., 2020. Machine learning based image processing technology application in bunch longitudinal phase information extraction. Physical Review Accelerators and Beams, 23(3), p.032805.

[9]. Sakkari, M. and Zaied, M., 2020. A convolutional deep self-organizing map feature extraction for machine learning. Multimedia Tools and Applications, 79, pp.19451-19470.

[10]. Kasongo, S.M. and Sun, Y., 2020. A deep learning method with wrapper based feature extraction for wireless intrusion detection system. Computers & Security, 92, p.101752.

[11]. Khan, M.A., Ashraf, I., Alhaisoni, M., Damaševičius, R., Scherer, R., Rehman, A. and Bukhari, S.A.C., 2020. Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists. Diagnostics, 10(8), p.565.

[12]. Devulapalli, S., Potti, A., Krishnan, R. and Khan, M.S., 2021. Experimental evaluation of unsupervised image retrieval application using hybrid feature extraction by integrating deep learning and handcrafted techniques. Materials Today: Proceedings.

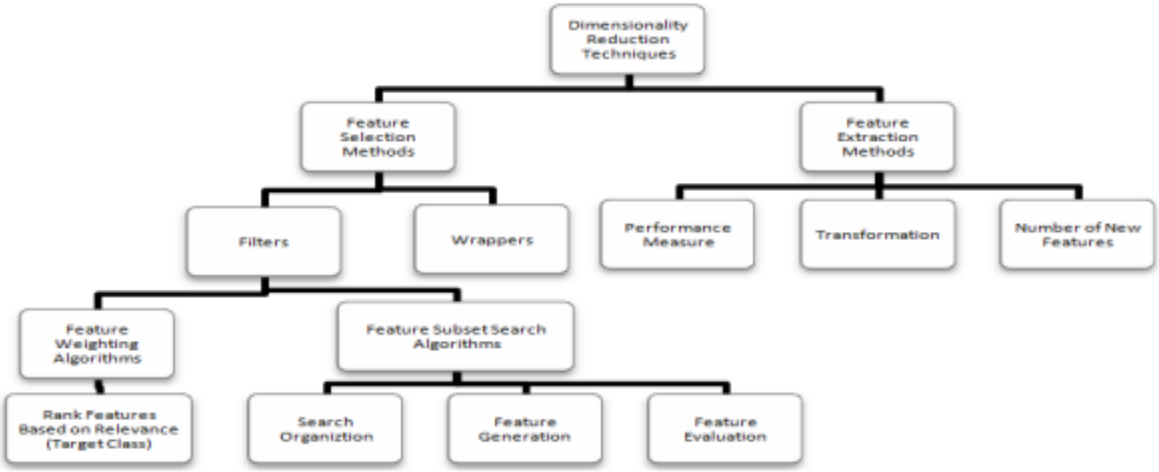

Fig 1: Hieratical Structure of Dimensionality Reduction Approaches

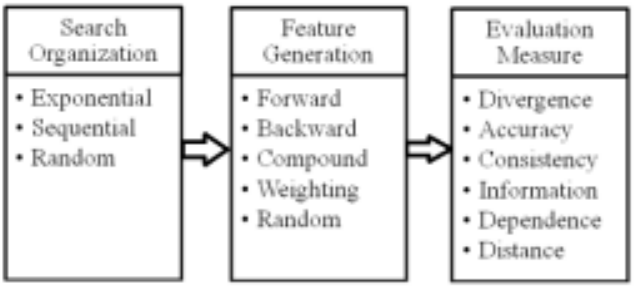

Fig 2: Characterization of Feature Selection Algorithms