A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning

Samina Khalid
Software Engineering Department
Bahria University Islamabad,
Pakistan
noshi_mir@yahoo.com

Tehmina Khalil
Software Engineering Department
Bahria University Islamabad,
Pakistan
Tehmina_khalil08@yahoo.com

Shamila Nasreen
Software Engineering Department
University of Engineering &
Technology Taxila, Pakistan
Shamila_nasreen131@yahoo.com

***Abstract*—Dimensionality reduction as a preprocessing step to machine learning is effective in removing irrelevant and redundant data, increasing learning accuracy, and improving result comprehensibility. However, the recent increase of dimensionality of data poses a severe challenge to many existing feature selection and feature extraction methods with respect to efficiency and effectiveness. In the field of machine learning and pattern recognition, dimensionality reduction is important area, where many approaches have been proposed. In this paper, some widely used feature selection and feature extraction techniques have analyzed with the purpose of how effectively these techniques can be used to achieve high performance of learning algorithms that ultimately improves predictive accuracy of classifier. An endeavor to analyze dimensionality reduction techniques briefly with the purpose to investigate strengths and weaknesses of some widely used dimensionality reduction methods is presented.**

Keywords- Age Related Macula Degeneration (AMD); Feature Selection; Feature Subset Selection; Feature Extraction/Transformation; FSA's; RELIEF, Correlation Based Method; PCA; ICA.

I. INTRODUCTION

Dimensionality reduction is wide spread preprocessing in high dimensional data analysis, visualization and modeling. One of the simplest ways to reduce dimensionality is by Feature Selection; one selects only those input dimensions that contain the relevant information for solving the particular problem. Feature Extraction is a more general method in which one tries to develop a transformation of the input space onto the low-dimensional subspace that preserves most of the relevant information [1]. Feature extraction and selection methods are used isolated or in combination with the aim to improve performance such as estimated accuracy, visualization and comprehensibility of learned knowledge [2]. Generally, features can be categorized as: relevant, irrelevant, or redundant. In feature selection process a subset from available features data are selected for the process of learning algorithm. The best subset is the one with least number of dimensions that most contribute to learning accuracy [3]. The advantage of feature selection is that important information related to a single feature is not lost but if a small set of features is required and original features are very diverse, there is chance of information lost as some of the feature must be omitted. On the other hand with dimensionality reduction also known as feature extraction, the size of feature space can be often decreased without losing information of the original feature space. A drawback of feature extraction is the fact that the linear combination of the original features are usually not interpretable and the information about how much an original feature contributes is often lost [4]. Studies demonstrate that a lot of efforts have been performed for devising best feature selection and feature extraction techniques and worth mentioning approaches are 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. In view of the substantial number of existing feature selection and feature extraction algorithms, the need arises to count on criteria that enable to adequately decide which algorithm to use in certain situations. A brief survey of these techniques based on literature review is performed to check suitability of different feature selection and feature extraction techniques in certain situation based on experiments that have been performed by researchers to analyze how these techniques helps to improve predictive accuracy of classification algorithm. In this study we make interested readers aware of different dimensionality reduction techniques.

The rest of the paper is organized as follow: section 2, consists on literature review of related surveys that have been performed to analyze performance of different dimensionality reduction techniques in medical field especially taking ophthalmology disease in consideration. Section 3, presents review of dimensionality reduction techniques to analyze how effectively these techniques can be used to achieve high performance of learning algorithms that ultimately improves predictive accuracy of classifier. In section 4 an endeavor to analyze dimensionality reduction techniques briefly with the purpose to investigate strengths and weaknesses of some widely used dimensionality reduction methods.

II. LITERATURE REVIEW

Dimensionality reduction techniques have become an obvious need in medical field (automated application). Today, a huge amount of data is generated in the medical domain. This includes the symptoms that a patient may have and also many medical test reports that may be generated. Feature is synonymous of input variables and attributes. In medical diagnosis example, the features can be symptoms, consist on a set of variables categorizing health status of a patient (e.g. in diabetic retinopathy symptoms of Dry or Wet Age related macular degeneration (AMD)). In this section literature review of some widely used feature selection and feature extraction methods in the detection and diagnosis of many eyes diseases for ophthalmologists (glaucoma, diabetic retinopathy and especially for automatic detection of age related macular degeneration) is presented. The main aim of this review is to make practitioners aware of the benefits, and in some cases even the necessity of applying dimensionality reduction techniques. To get benefit from dimensionality reduction techniques for the purpose of maximizing accuracy of learning algorithm, there is need to have awareness of various advantages of these techniques. L. Ladha et al in [3] have been offered following advantages of feature selection:

- It reduces the dimensionality of the feature space, to limit storage requirements and increase algorithm speed.
- It removes the redundant, irrelevant or noisy data.
- The immediate effects for data analysis tasks are speeding up the running time of the learning algorithms.
- Improving the data quality.
- Increasing the accuracy of the resulting model.
- Feature set reduction, to save resources in the next round of data collection or during utilization.
- Performance improvement, to gain in predictive accuracy.
- Data understanding to gain knowledge about the process that generated the data or simply visualizes the data.

P. Soliz and et al [6] have proposed an approach for extracting image-based features for classifying AMD in digital retinal image. 100 images have been classified by an ophthalmologist into 12 categories based on the visual characteristics of the disease. Independent components analysis (ICA) has been used to extract features and used input to classifier. It has been shown that ICA can robustly detect and characterized features in funds images, and extract implicitly the mathematical features from each image to define the phenotype. M. Pechenizky [7] has been analyzed the effects of class noise (misclassification or mislabeling) on supervised learning in medical domains. A review of related work on learning from noise data has been discussed and proposed to use feature extraction as a pre-processing step to diminish the effect of class noise on the learning process. The filtering techniques handle noise explicitly. Many filtering approaches have been summarized that have been acknowledged useful by researchers. However, the same researchers have recognized some practical difficulties with filtering approaches. One concern is that it is difficult to differentiate noise from exception (outliers) without the help of an expert. Another concern is that a filtering technique can use an expected level of noise as an input parameter, and this value is rarely known for a particular datasets. Feature extraction (using PCA) techniques fits better for noise-tolerate techniques as it helps to avoid over fitting implicitly within learning techniques. Feature extraction techniques before undertaking supervised learning indeed enables decreasing the negative effect of the presence of mislabeled instances in the data. In [8] diabetic detection method using ANN and a feature set formed by adopting singular value decomposition (SVD) and Principle Component Analysis (PCA) has been proposed. Experimental results show that ANN-SVD+PCA composition is a reliable means of diabetes detection with less computational cost and high accuracy. Feature extraction methods found much more suitable for automated detection of ophthalmologists diseases than feature selection methods because of noisy data. Because most of the datasets related to bio medical contain noisy data instead of irrelevant or redundant data.

III. DIMENSIONALITY REDUCTION APPROACHES

High dimensional data is problematic for classification algorithms due to high computational cost and memory usage [4]. There are two dimensionality reduction techniques know as feature extraction (also known as dimensionality reduction explicitly or feature transformation) and feature selection (FS). The advantage of FS is that no information about importance of a single feature is lost but if a small set of features is required and original features are very diverse, information can be lost as some of the feature must be omitted during feature subset selection process. Whereas, in feature extraction size of the feature space can often be decreased without losing a lot of information of the original feature space. The choice between feature extraction and feature selection methods depends on specific data type of domain of application.

3.1 Feature Selection

High dimensional data consists on features that can be irrelevant, misleading, or redundant which increase search space size resulting in difficulty to process data further thus not contributing to the learning process. Feature subset selection is the process of selecting best features among all the features that are useful to discriminate classes. Feature selection algorithm (FSA) is a computational model that is provoked by a certain definition of relevance. L. Ladha et al [3] have been presented an empirical comparison of different feature selection algorithms. In general feature selection is

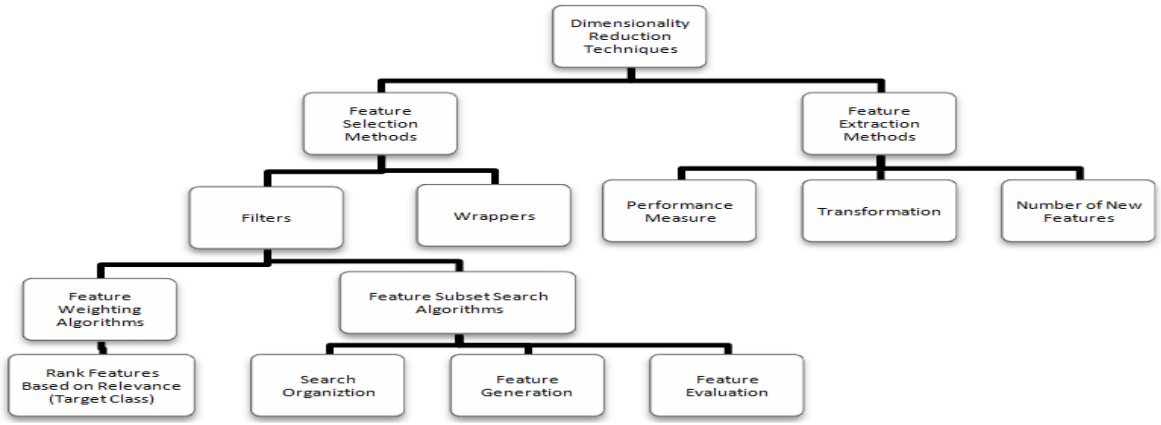

Fig 1: Hierarical Structure of Dimensionality Reduction Approcaches

known as a search problem according to some evaluation criteria. Feature selection algorithms can be characterized by (i) search organization: three types of search is possible exponential, sequential, or random. (ii) Generation of successors (subset): five different operators can be considered to generate successor that are; Forward, Backward, Compound, Weighted, and Random. (iii) Evaluation Measure: evaluation of successors can be measure through Probability of Error, Divergence, Dependence, Interclass Distance, Information or Uncertainty and Consistency Evaluation as shown in Figure 2. Feature selection methods can be distinguished into three categories: filters, wrappers, and embedded/hybrid method. Wrappers methods perform well than filter methods because feature selection process is optimized for the classifier to be used. However, wrapper methods have expensive to be used for large feature space because of high computational cost and each feature set must be evaluated with the trained classifier that ultimately make feature selection process slow. Filter methods have low computational cost and faster but with inefficient reliability in classification as compared to wrapper methods and better suitable for high dimensional data sets. Hybrid/embedded methods are recently developed which utilize advantages of both filters and wrappers approaches. A hybrid approach uses both an independent test and performance evaluation function of the feature subset [10]. Filters methods can be further categorized into two groups, namely feature weighting algorithms and subset search algorithms as shown in Figure 1. Feature weighting algorithms assign weights to features individually and rank them based on their relevance to the target concept [11]. A well known algorithm that relies on relevance evaluation is Relief (discussed in section 4).

3.2 Feature Extraction/Transformation

Features extraction performs some transformation of original features to generate other features that are more significant. Brian Ripley [8] defined feature extraction as follow: "Feature extraction is generally used to mean the construction of linear combinations αTx of continuous features which have good discriminatory power between classes". An important problem in Neural Networks research and other disciplines like Artificial Intelligence is facing in finding a suitable representation of multivariate data. Features extraction can be used in this context to reduce complexity and give a simple representation of data representing each variable in feature space as a linear combination of original input variable. The most popular and widely used feature extraction approach is Principle Component Analysis (PCA) introduced by Karl. Many variants of PCA have been proposed. PCA is a simple non-parametric method used to extract the most relevant information from a set of redundant or noisy data. PCA is a linear transformation of data that minimizes the redundancy (measured through covariance) and maximizes the information (measured through the variance) [12].

In [4] to investigate the relationship between various dimensionality reduction methods (including feature subset selection using information gain (IG) and wrapper methods, and feature extraction with deferent flavors of PCA methods) and effects of these methods on classification performance have been empirically tested on two different types of data sets (e-mail data and drug discovery data). Results have shown that feature extraction (transformation) using PCA is highly sensitive to the type of data. Feature selection method Wrapper shows reasonable effect on classification accuracy than IG for both types of data. The experimental results underline the importance of a dimensionality reduction process. Wrappers methods for feature selection tend to produce smallest features subsets with very competitive classification accuracy as compare to feature extraction methods. However, wrappers are much more computationally expensive than the feature extraction methods [13, 14]. Veerabhadrappa and L. Rangarajan in [10, 15] proposed bi-level dimensionality reduction methods that have integrated feature selection and feature extraction methods with the aim to improve classification performance. They have proposed two approaches, in first

level of dimensionality reduction; features are selected based on mutual correlation. In second level selected features are used to extract features using PCA and LPP. Proposed method applied to several standard datasets to evaluate its performance. The results obtained shows that proposed system outperform over single level dimensionality reduction techniques.

IV. ANALYSIS OF DIFFERENT FEATURE SELECTION AND FEATURE EXTRACTION TECHNIQUES

4.1 Feature Selection Algorithms (FSA)

An overview of some basic FSA's with their limitations has been discussed by L. Ladha et al [3]. *Chi- squared* is the most common used statistical test that measures divergence from the distribution expected if one assumes the feature occurrence is actually independent of the class values. *Euclidian Distance* examines the root of square differences between coordinates of a pair of objects. The advantage of this method is that the distance is not affected by the addition of new objects to the analysis, which may be outliers. However, Euclidian distance can be greatly affected by differences in scale among the dimension from which the distance is computed. The *t-test* assesses whether the average of two groups are statistically different from each other. This analysis is esteemed whenever there is need to compare the average of two groups, and especially suitable as the analysis for the posttest-only two-group randomized experimental design. *Information Gain (IG)* measures the increase in entropy when the feature is given vs. absent. This is the application of more general techniques, the measurement of informational entropy, to the problem of deciding how important a feature is within feature space. *Correlation-Based Feature Selection (CFS)* searches feature subset according to the degree of redundancy among the features. The evaluation process aims to find subsets of features that are individually highly correlated with the class but have low inter-correlation. Relevance of group of features grows with the correlation between features and class, and decreases with growing inter-correlation. CFS is used to determine the best feature subset and is usually combined with search strategies such as forward selection, backward elimination, bi-directional search, best first search and genetic search.

In [11] Lei Yo et al, introduced a novel concept, predominant correlation, and proposed a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis. *Sequential Forward Selection (SFS)* is simplest greedy search algorithm. SFS performs best when the optimal subset has a small number of features. The main disadvantage of SFS is that it is unable to remove features that become obsolete after the addition of other features. *Sequential Backward Elimination (SBE)* works in opposite direction of SFS. SBE work best when the feature subset has a large number of features. The main limitation of SBE is its inability to reevaluate the usefulness of a feature after it has been discarded. *Plus-L Minus-R Selection (LRS)* is a generalization of SFS and SBE. It attempts to compensate for the weaknesses of SFS and SBE with some backtracking capabilities. Its main limitation is the lack of theory to help predict the optimal values of L and R. the problem with individual feature selection algorithms is that they only capture the relevance of features to the target concept and avoid redundancy among features. Empirical evidence from features selection literature shows that, along with irrelevant features, redundant features also affects the speed and accuracy of learning algorithms and thus should be eliminated as well. Therefore in the context of feature selection for high dimensional data where there may exist many redundant features, pure relevance based feature weighting algorithms do not meet the need of feature selection very well [11].

Characteristics of Feature Subset Selection

Subset search algorithms search through candidate feature subsets guided by a certain evaluation measure, which captures the goodness of each subset. In [16] performance of some feature selection algorithms have been analyzed using various datasets from public domain. Number of reduced features and their effect on learning performance with some widely used methods have been measured, then evaluated and compared. Feature selection method should choose the best feature subset from feature space to describe target conceptions of learning process. There must be considering following aspects in the process of feature selection: 1. Starting Point, 2. Search Strategy, 3. Subset Evaluation, 4. Stopping Criteria. On the basis of these aspects comparative analysis of feature selection techniques is shown in Table 1. We characterized feature selection techniques in a way to give an overview of comparative analysis about search organization, feature generation, and evaluation measure that each feature selection technique implies which can guide the choice for a technique suited to the goals and resources of practitioners in the field. Nine feature selection methods have been discussed in [16]. *mRMR* (Minimal Redundancy and Maximal Relevance) uses mutual information (MI) of two random variables. MI is a quantity that measures the mutual dependency of the two variables.

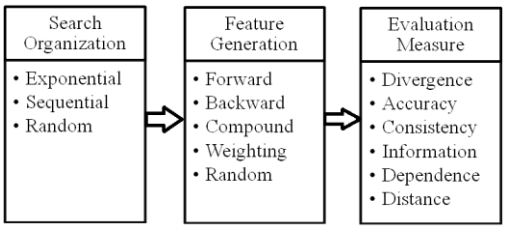

Fig 2: Charachertization of Feature Selection Algorithms

This method uses MI between a feature and a class as relevance of the feature for the class, and MI between features as redundancy of each feature. *I-RELIEF* is well known weighting (ranking) method that measures the

relevance of features in the neighbored around target samples. Relief finds the nearest sample in feature space of same category called hit sample, then measures the distance between target and hit samples. It also finds nearest sample of opposite category called miss sample and then does same work. Difference between those measured distances has used as the weight of the target feature. This basic algorithm was extracted into several variants. *I-RELIEF* approach reduces the biasness of original RELIEF method. *Conditional Mutual Information Maximization (CMIM)* selects a feature subset that has maximum relevance to the target class by using conditional mutual information. CMIM requires both the feature values and output classes be binary.

Correlation Coefficient approach evaluates how well an individual feature contributes to the separation of classes. Ranking criteria is used to rank all features using their mean and standard deviation for all the samples of both classes. *Between-Within Ratio (BW-ratio)* uses the ratio of between group to within group sums of squares for each feature, and select feature with maximum BW-ratio. A single variable can be considered irrelevant based on its correlation to the class, but when combines with other features in the feature space, it becomes very relevant [17]. *INTERACT* methods considers feature interaction. This approach finds interacting features by backward elimination with measurement of consistency contribution. C-consistency of a feature is an indicator about how significantly the elimination of a feature will affect consistency e.g. C-consistency of irrelevant feature will be zero. *Genetic Algorithm* is a randomized approach. GA consist on particular class of evolutionary algorithms that makes use of techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover. Each feature set is represented by a binary string in feature selection problems [18]. *Recursive Feature Elimination (SVM-RFE)* is a wrapper method which performs backward elimination. SVM-REF finds the *m* features which lead to the largest margin of class separation, and uses the weight vector *w* as a ranking criterion [19]. *Prediction Analysis of Microarray (PAM)* is a statistical method for class prediction using gene expression data using shrunken centroid. The method of nearest shrunken centroid identifies subsets of genes that the best characterized the class [20]. An experimental analysis of above mentioned feature subset selection techniques has been presented in [16]. Seven data sets including lung cancer, leukemia including five other data sets from UCI machine learning have been taken and preprocessed for discrete feature selection.

Table1: Comparative Analysis of Feature Selection Algorithms

Methods	Individual / Subset Feature	Starting Point	Search Strategy	Subset Generation	Subset Evaluation	Stopping Criteria	Used to Eliminate
Correlation Coefficient	Individual	Random Number of Features	Sequential	Forward Selection	Divergence (variance)	Ranking	Irrelevant Features
BW-ration	Individual	Full Feature Set	Sequential	-	Divergence (variance)	Ranking	Irrelevant Features
PAM	Individual	Random Number of Features	Sequential	Weighted	Distance/Information	Ranking	Irrelevant Features
mRmR	Subset	Random Number of Features	Random	Forward Selection	Mutual dependence/information	Ranking	Redundant Features / Irrelevant Features
I-RELIEF	Subset	Random Number of Features	Random	Weighted	Distance	Ranking	Irrelevant Features
CMIM	Subset	Full Feature Set	Sequential	Forward Selection	Conditional Mutual Information /	Relevance	Irrelevant Features
INTERACT	Subset	Full Feature Set	Sequential	Backward Elimination	Consistency	Relevance	Irrelevant Features
Genetic Algorithm	Subset	Full Feature Set	Random	Weighted	Consistency (cosine)	Ranking	Redundant Features / Noise
SVM-REF	Subset	Full Feature Set	Sequential	Backward Elimination/Weighted	Information	Ranking	Irrelevant Feature

To evaluate feature subset selection techniques two learning algorithms Naiv Bayes and LIBSVM algorithms have been used. Almost all above mentioned techniques performed well with reduced feature subset. *mRMR* method performed better than all other methods for most data sets. On the other hand *I-RELIEF* method performed poor for discrete data. *GA* also

produced poor results for microarray. From this experimental analysis it has concluded that feature selection methods that handle elimination of both redundant and irrelevant features at once are much more robust and beneficial for learning process as compared to methods that discretely handle feature redundancy and/or irrelevant features.

4.2 Feature Extraction/Transformation Methods

It is important for subsequent analysis of the data; either it is pattern recognition, de-noising, data compression, visualization, or anything else that the data is represented in a manner that facilitates the analysis. Several principle methods have been developed to find a suitable transformation. *Independent Component Analysis (ICA)* is a linear transformation method, in which the desired representation is the one that minimize the statistical dependence of the components of the representation. The use of ICA for feature extraction is motivated by results in neurosciences that suggest that the similar principle of redundancy reduction explains some aspects of the early processing of sensory data by the brain. ICA has also applications in exploratory data analysis in the same way as the closely related method of projection pursuit. The use of feature extraction is motivated by the theory of redundancy reduction [21]. There are two main families of ICA algorithms. Some algorithms are rooted in the minimization of mutual information; others take root in the maximization of non gaussianity. Mutual information can be seen as the reduction of uncertainty regarding variable X after the observation of Y. Therefore by having an algorithm that seeks to minimize mutual information, we are searching for components that are maximal independent. Another way to estimate the independent component is by focusing on non gaussianity. One way to extract the components is by forcing each of them to be as far from the normal distribution as possible. Usually, to perform ICA five conditions must be met: 1 – the source signals must be statistically independent; 2- the number of source signals must be equal the number of mixed observed signals and mixtures must be linearly independent from each other; 3- the model must be noise free; 4- data must be centered; 5- the source signals must not have a Gaussian probability density function (pdf), except for one signal source that can be Gaussian [22].

Principle Component Analysis (PCA) consists into an orthogonal transformation to convert samples belonging to correlated variables into samples of linearly uncorrelated features. The new features are called principle components and they are less or equal to the initial variables. PCA is an unsupervised technique and as such does not include label information of the data. If data are normally distributed then principle components are independent. The main reason for the use of PCA concerns the fact that PCA is a simple non-parametric method used to extract the most relevant information from a set of redundant or noisy data.

Table 2: Comparison of FSA's

Methods	Time Complexity	Data Type	Application
mRmR	-	Discrete / Continues	Microarray Gene Expressions
I-RELIEF	O(Poly(N))	Discrete / Continues / Nominal	Protein Folding and Weather Prediction
CMIM	$O(N^3)$	Boolean	Image Classification
Correlation Coefficient	-	Discrete / Continues	-
BW-ration		Discrete / Continues	-
INTERACT	$O(N^2M)$	-	-
Genetic Algorithm	-	-	Pattern Recognition, Machine Learning, Neural Networks, Combinatorial Optimization, Hyper Spectral Data
SVM-REF	-	Discrete / Continues	Microarray Gene Expressions
PAM	-	-	Microarray Gene Expressions

PCA reduces the number of original variables by eliminating the last principle components that do not contribute significantly to the observed variability. PCA is a linear transformation of data that minimize the redundancy (measured through covariance) and maximize information (Measured through variance). Principle components (PC) are new variables with two properties: 1) each PC is a linear combination of the original variables; 2) the PC's are uncorrelated to each other and also the redundant information is removed [12]. Main application areas of PCA include data compression, image analysis, visualization, pattern recognition, regression, and time series prediction. PCA suffers from some limitations;

1. It assumes that the relationships between variables are linear.
2. Its interpretation is only sensible if all of the variables are assumed to be scaled at the numeric level.
3. It lacks a probabilistic model structure which is important in many contexts such as mixture modeling and Bayesian decision.

An alternative method to circumvent first and second limitations has been proposed by Guttman (1941) named *Nonlinear Principle Components Analysis*. This method has same objective as of PCA but is suitable for variables of mixed measurement level (nominal, ordinal and numeric). In nonlinear PCA all variables are viewed as categorical, and every distinct value of a variable is referred to as a

category that's why also called categorical PCA [23]. Nominal variables cannot be analyzed by standard PCA. Ordinal variables are also referred to as categorical. Such variables consist of ordered categories, such as the values on a rating scale for e.g. Likert type scale. The crucial difference is that in linear PCA the measures variables are directly analyzed, whereas in nonlinear PCA the measured variables are quantified during analysis. *Probabilistic Principle Component Analysis (PPCA)* overcomes third limitation by letting the noise component possess an isotropic structure; the PCA is implicitly embedded in a parameter learning stage for this model using the maximum likelihood estimation method. An efficient expectation/maximization (EM) algorithm is also developed to iteratively learn the parameters. *Kernel Principle Component Analysis (KPCA)* overcomes the first limitation by using a kernel trick. The essential idea of KPCA is avoid the direct evaluation of the required dot product in a high dimensional feature space using the kernel function. Therefore, no explicit nonlinear function projecting the data from the original space to the feature space is needed. In [24] an approach to analyze kernel principle component in a probabilistic manner has been proposed called probabilistic kernel principle component analysis (PKPCA) that naturally combines PPCA and KPCA to overcome limitations of PCA.

V. DISCUSION

Feature extraction methods found much more suitable for automated detection of ophthalmologists diseases than feature selection methods because of noisy data. Because most of the datasets related to bio medical contain noisy data instead of irrelevant or redundant data. Feature selection is used in many application areas as a tool to remove irrelevant and/or redundant features. There is no single feature selection method that can be applied to all applications. Some methods used to eliminate irrelevant features but avoid redundant features. Pure relevance based features weighting algorithm do not meet the need of feature selection very well. Subset search algorithms search through candidate feature subsets guided by a certain evaluation measure which captures the goodness of each subset. Some existing evaluation measures that have been shown effective in removing both irrelevant and redundant features include the consistency measure and correlation measure. Experiments show that in order to find best feature subset, the number of iterations required is mostly at least quadratic to the number of features. Therefore, with quadratic or higher time complexity in term of dimensionality, existing subset search algorithms do not have strong scalability to deal with high dimensional data. Feature selection methods broadly categorized into filters and wrappers. Wrapper methods generally result in better performance than filter methods because the feature selection process is optimized for the classification algorithm to be used. However, they are generally far too expensive to be used if the number of features is large because each feature set considered must be evaluated with the trained classifier. Filters methods are much faster than wrapper methods and therefore are better suited to high dimensional data sets. To combine the advantages of both models, algorithms in a hybrid model have recently been proposed to deal with high dimensional data. Table 1: Presents a comparative analysis of different feature selection algorithms showing that most of the methods handle redundant feature elimination or irrelevant features separately. And very few methods handle noisy data. Feature extraction methods have been proposed as a preprocessing step to diminish the effect of class noise on the learning process. Studies show that classification accuracy achieved with different feature reduction strategies is highly sensitive to the type of data. Feature selection methods that handle elimination of both redundant and irrelevant features at once are much more robust and beneficial for learning process as compared to methods that discretely handle feature redundancy and/or irrelevant features.

VI. CONCLUSION

A survey of feature selection and extraction is proposed. The objective of both methods concerns the reduction of feature space in order to improve data analysis. This aspect becomes more important when real world datasets are considered, which can contain hundreds or thousands features. The main difference between feature selection and extraction is that the first performs the reduction by selecting a subset of features without transforming them, while feature extraction reduces dimensionality by computing a transformation of the original features to create other features that should be more significant. Traditional methods and their recent enhancements as well as some interesting applications concerning feature selection are presented in table 2. Feature selection improves knowledge of the process under consideration, as it points out the features that mostly affect the considered phenomenon. Moreover the computation time of the adopted learning machine and its accuracy need to be considered as they are crucial in machine and data mining applications.

REFERENCES

[1] N. Chumerin and V. Hulle, M. M, "Comparison of Two Feature Extraction Methods Based on Maximization of Mutual Information" In: Proceedings of the 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, pp. 343–348, 2006.

[2] H. Motoda and H. Liu, "Feature selection, extraction and construction" In: Towards the Foundation of Data Mining Workshop, Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2002), Taipei, Taiwan, pp. 67–72, 2002.

[3] L. Ladla and T. Deepa, " Feature Selection Methods And Algorithms", International Journal on Computer Science and Engineering (IJCSE), vol.3(5), pp. 1787-1797, 2011.

[4] A. G. K. Janecek and G. F. Gansterer et al, "On the Relationship between Feature Selection and Classification Accuracy", In: Proceeding of New Challenges for Feature Selection, pp. 40-105, 2008.

[5] M. Dash and h. Liu, "Feature Selection for Classification", Intelligent Data Analysis, vol. 1, pp. 131-156, 1997.

[6] P. Soliz et al, "Independent Component Analysis for Vision-inspired Classification of Retinal Images with Age-related Macular Degeneration", In: proceeding of IEEE Int'l. Conference on image processing SSIAI, pp. 65-68, 2008.

[7] A. Deka and K. Kumar Sarma, "SVD and PCA Feature for ANN Based Detection of Diabetes Using Retinopathy", In: Proceedings of the CUBE International Information Technology Conference, pp.38-44, 2012.

[8] Y. Zheng et al, "An Automated Drusen Detection System for Classifying Age-Related Macular Degeneration with Color Fundus Twentieth International Conference on Machine Learning (ICML), 2003.

[12] S. Cateni, et al, "Variable Selection and Feature Extraction through Artificial Intelligence Techniques", Multivariate Analysis in Management, Engineering and the Science, chapter 6, pp.103-118, 2012.

[13] T. Howley and M. G. Madden et al, " The Effect of Principle Components Analysis on Machine Learning Accuracy with High Dimensional Spectral Data" Knowledge Based Systems, vol. 19(5), pp. 363-370, 2006.

[14] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection" *Journal of Machine Learning Research*, vol.3, pp. 1157-1182, 2003.

[15] Veerabhadrappa, L. Rangarajan, "Multilevel Dimensionality Reduction Methods Using Feature Selection and Feature Extraction", International Journal of Artificial Intelligence and Applications, vol. 1(4), pp. 54-58, 2010.

[16] C. Yun and J. Yang, "Experimental Comparison of Feature Subset Selection Methods", In: Seventh IEEE International Conference on Data Mining- workshop, pp. 367-372, 2007.

[17] Z. Zhao, H. Liu, "Searching for Interacting Features", Proceedings of International Joint Conference on Artificial Intelligence, pp. 1156-1161, 2007.

[18] S. Lei, "A Feature Selection Method Based on Information Gain and Genetic Algorithm", In: International Conference on Computer Sciences and Electronics Engineering, pp. 355-358, 2012.

[19] I. Guyon et al, "Gene Selection for Cancer Classification using Support Vector Machines", Machine Learning, Vol. 46(1-3), pp. 389-422, 2002.

Photographs", In: IEEE 10th International Symposium on Biomedical Imaging, pp.1440-1443, 2013.

[9] S. Rajarajeswari and K. Somasundaram, " An Empirical Study of Feature Selection For Data Classification", International Journal of Advanced Computer Research, vol.2(3) issue-5, pp. 111-115, 2012.

[10] Veerabhadrappa, L. Rangarajan, "Bi-level Dimensionality Reduction Methods Using Feature Selection and Feature Extraction", International Journal of Computer Applications, vol. 4(2), pp. 33-38, 2010.

[11] L. Yu, and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", In: Proceeding of the

[20] R.Tibshirani, T.Hastie, B.Narasimhan and G.Chu, "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression", In; Proceedings of the National Academy of Sciences of the United States of America, Vol. 99(10), pp. 6567-6572, 2002.

[21] A. Hyyarinen, "Survey on Independent Component Analysis", Neural Computing Surveys, vol. 2, pp.94–128, 1999.

[22] D. Langlois et al, "An Introduction to Independent Component Analysis: InfoMax and FastICA Algorithms", Tutorials in Quantitative Methods for Psychology, vol. 6(1), pp. 31-38, 2010.

[23] Linting et al, "Nonlinear Principle Components Analysis: Introduction and Application", Psychological Methods, 2007.

[24] S. Zhou, "Probabilistic analysis of Kernel Principle Components: Mixture Modeling, and Classification", Draft, pp. 1-26, 2003.

[25] M. Pechenizkiy et al, "Class Noise and Supervised Learning Learning in Medical Domain: The Effect of Feature Extraction" In: Proceeding of the 19th IEEE Symposium on Computer-Based Medical System (CBMS'06), 2006.

[26] C. Ding and H. Ping, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data", Journal of Bioinformatics and Computational Biology, vol. 3(9), pp. 185-205, 2005.

[27] F. Fleuret, "Fast Binary Selection with Conditional Mutual Information", Journal of Machine Learning, pp. 1531-1555, 2004.

[28] I. Guyon, "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research 3, pp. 1157-1182, 2003.

[29] M. Klassen and N. Kim, "Nearest shrunken centroid as feature selection of microarray data," In: Proceedings of Computers and Their Applications, pp. 227-232, 2009.