Dimension Reduction: A Review

Mohini D. Patil
Post Graduate Student,
K. K. W.I.E.E.R, Nasik
(Affiliated to University of Pune)

Shirish S. Sane, Ph.D
Head of Computer Engineering Department,
K. K. W.I.E.E.R, Nasik
(Affiliated to University of Pune)

ABSTRACT
Dimension reduction is the process of keeping only those dimensions in a dataset which are important from the point of view of problem at hand and discarding of the others. This helps to design easily computable algorithms and to increase the performance of classifiers. It has gained importance as a preprocessing step in knowledge discovery and data mining especially in the fields of pattern matching, machine learning, bioinformatics and genetics which involve datasets having large number of dimensions. There are two basic strategies used for reduction of the dimensions; feature selection and feature extraction. Feature selection techniques focus on selecting some of the important features from all the features while feature extraction techniques are based on generating new features by making use of entire information present in the original dataset. Recently some work has also focused upon combining both the strategies to club their advantages. This paper studies the basic strategies for dimension reduction, the different techniques proposed in literature for reducing the dimensions and also about the measures used by them. Finally it also discusses about the combined approach and concluding remarks are given.

General Terms
Data Mining, Preprocessing

Keywords
Dimension reduction, Feature Selection, Feature Extraction.

1. INTRODUCTION
Dimension Reduction is the process of converting an *n-dimensional* space problem to a *p-dimensional* space problem where $p < n$. This section gives a brief overview of it.

1.1 Need
The high dimensional data refers to the data having large number of attributes/features. Such data is prominent in areas of engineering, biometrics, machine learning, data mining etc. The large number of dimensions present in the dataset poses problems for the classifier and may result in reduction of accuracy. Thus to cope up with this problem and increase the accuracy of classifier there is need to remove the insignificant and irrelevant features or to extract new features which will be richer in content and would be able to describe the original dataset without any loss. Dimension reduction is an important preprocessing step to handle this problem [1].

1.2 Goal
The goal of dimension reduction is to represent the dataset with minimum number of attributes such that they will result in the same or comparable performance as that obtained using all the dimensions [1].

1.3 Issues
While reducing the dimensions of a dataset following issues need to be considered[3][4][5][7][9]

- Measures Used: Different measures like dependency, relevance, significance, mutual information etc can be used for reducing the dimensions. They must be chosen correctly depending upon the problem at hand.
- Incomplete and Missing data: The real world datasets mostly have some of the values that are missing or are incomplete. Rough set theory has been noted in the literature for this purpose.
- Discrete or continuous valued data: The dataset may also contain data in various forms like discrete, continuous etc. Some of the techniques can handle the continuous valued data directly while some require discretization. Hence this must be considered.
- Vagueness of the data: The real world dataset may also contain data which is fuzzy. The use of fuzzy set and fuzzy rough set theory has been proved to be a good tool to deal with this.

1.4 Organization of the paper
Section 1 gives an introduction to the concept of dimension reduction. Section 2 describes the feature selection strategies and techniques. Section 3 focuses on the feature extraction techniques. Section 4 briefly describes the simultaneous feature selection and extraction and finally the paper is concluded in Section 5.

1.5 Basic Terminologies
The following terms are used throughout the paper:

- Dimensions: It refers to the number of features or attributes of the dataset. The words dimension/ attributes /features are used interchangeably.
- Decision Attributes: They refer to the attributes which have decision values/class labels. For example in case of a cancer data set the final result of the sample as positive or negative is a class label and the attribute describing it is the decision attribute.
- Conditional Attributes: The attributes other than the decision attributes are called as conditional attributes.
- Criterion function: It is a mathematical function used for finding the importance of a dimension.

2. FEATURE SELECTION
Feature selection is the strategy of selecting only those features from all, which are relevant, significant and important from the point of view of classification or clustering. The less important features are discarded by it.

This section first discusses about the basic algorithmic strategies used for feature selection, and then a discussion on some of the feature selection techniques as given in the literature is done along with their features and shortfalls.

2.1 Basic Algorithmic Strategies

The basic strategies used for selecting the attributes (features) are listed below [1][2]:

1. Exhaustive Search: This technique tries out all the combinations of the features for the required number of features and selects that combination which maximizes the criterion function. The criterion function may be based on relevance, dependency significance etc. For example if number of features is n and required features are k then C_k^ncombinations need to be carried out.
   Drawback: It is not suitable where the number of features is too large as it becomes cumbersome for computations.
2. Branch and Bound: It forms a tree where the root pertains to choosing all the features. The children at next level of root consist of combination of features by removing one feature. From each of these children, new nodes are formed where another feature has been removed and so on. A leaf of the tree represents combination of features. Once a leaf node is reached, the criterion function is evaluated; value is found and stored as a bound. While evaluation of future branches if criterion function's values go below bound, then that branch is not evaluated. When another leaf is reached if its criteria value is greater than bound then it is updated and that combination of features is stored as the best so far. For example: Refer the example in Fig 1, The strategy starts with root node having all the four features f1,f2,f3,f4. At the next level a combination of three out of four features is carried out and the criteria functions value(J) is evaluated. The first node is evaluated fully to find out the bound. The further branches from second node onwards are evaluated only if the value of J does not fall below the bound. Hence the node E is not evaluated as already node C had evaluated to value 28 which is greater than that of value at E which is 26. Thus after exploring the node C,it gives maximum value and the combination f2,f4 is chosen as the best reduct.
   Drawback: The subset generated may not be the optimal one as some of the nodes are not expanded. But this drawback is overcome by relaxed version of branch and bound where the nodes are allowed to be evaluated if their value is within some margin. However, the selection of margin must be done carefully.

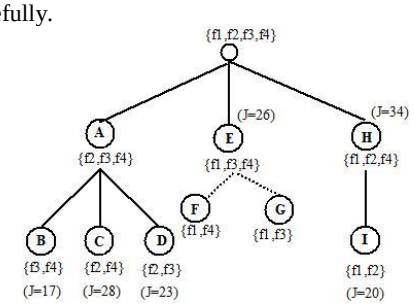

Figure 1:Example1

3. Selecting Individual Features: Every feature is evaluated individually and after evaluating all of them, the best f features are selected. Computationally it is fast as no combinations need to be done.
   Drawback: The subset chosen is not likely to give good results since dependencies between features are not taken into account.
4. Sequential Methods: Here, the features are sequentially added or removed at each step based on the criterion function. There are two basic techniques which are listed below.
   i) *Sequential Forward Search:* This search starts with an empty set and keeps on adding features. It adds one feature at a time. The feature being added is the most significant or relevant one. Once a feature gets added, then it cannot be removed.
   For example: Consider training data as

   Class 1:1-4, 5-8, 1-4, 9-12
   Pattern1: (5,7,3,12)
   Pattern2 :(1,4,2,10)
   Pattern3: (3,8,1,8)
   Pattern4: (2,4,1,11)

   Class 2:5-8,1-4,9-12,5-8
   Pattern5: (6,2,10,5)
   Pattern6: (4,3,8,5)
   Pattern7: (7,1,12,9)
   Pattern8: (8,4,11,9)

   Consider validation data as

   Class 1: Pattern1(3,7,3,9) Pattern2(4,6,1,12)
   Pattern3(2,5,3,10) Pattern4(3,7,1,12)
   Class 2: Pattern5(7,2,12,6) Pattern6(5,1,10,8)
   Pattern7(6,3,11,6) Pattern8(7,4,9,7)
   Here if we classify the training set based on validation and using single feature at a time then :
   Feature 1 classifies 6 out of 8 patterns
   Feature 2 classifies 6 out of 8 patterns
   Feature 3 classifies 7 out of 8 patterns
   Feature 4 classifies 5 out of 8 patterns.
   As feature 3 gives maximum number of correct classifications it is selected. The process continues with the next step having combinations of 2 features and so on till the required numbers of features are selected.
   ii) *Sequential Backward Search*: It starts with all the features selected and in every iteration a feature is discarded. The feature being removed is the one which is most insignificant or irrelevant. Also once a feature is discarded, it cannot be added again.
   For example: Considering the same example as that of forward search, the sequential backward search will start with all the features selected (f1,f2,f3,f4) and then it will remove the feature which misclassified the data to maximum extent. In the above example it is f4,so it will be removed. This process will continue till the required features remain in the set.

   Drawbacks of the above two techniques is that of nesting effect.

   iii) *Hybrid Approach:* It adds some features and removes some of them in each iteration. Features

being added are the most significant ones and the ones being removed are the least significant ones.

Drawback: The number of features to be added and deleted in every iteration must be chosen properly.

2.2 Feature Selection Algorithms

Few of the feature selection algorithms are discussed below:

1) P.Maji et.al[3] have used the criteria of relevance and significance for selecting the important genes from the microarray gene expression data for diagnosis.
The measures used for calculating the total relevance and total significance are given as

$$Relevance = \tau_{relev} = \sum \hat{f}\ (A_i, D) \quad (1)$$

$$Significance = \tau_{signf} = \sum_{A_i \neq A_j \in S} \tilde{f}\ (A_i, A_j) \quad (2)$$

Here $\hat{f}\ (A_i, D)$ is the relevance of individual feature A_i with respect to decision label D and $\tilde{f}\ (A_i, A_j)$ is the significance of feature A_j with respect to A_i and they are calculated using rough set theory [3][5]

Algorithm

Input: $C = \{A_1, A_2 \ldots\ldots A_i \ldots\ldots A_m\}$ is the set of m features of given data set

Output: S is the set of selected features

Steps:

i) Initialize $C = \{A_1, A_2 \ldots\ldots A_i \ldots\ldots A_m\}, S \leftarrow \emptyset$

ii) Calculate the relevance of each feature $A_i \in C$

iii) Select the feature A_i as most relevant attribute. Add it to S and remove it from C.

iv) Repeat the following (v,vi) steps till desired number of genes are selected.

v) Calculate the significance of each of the remaining features in C with respect to the selected genes in S. Remove the feature from C if it has zero significance with respect to any of the selected genes.

v) From remaining features select the one which maximizes the sum of relevance and significance given in eq(1) and eq(2).

Drawback: The relevance and significance are calculated using rough set theory and hence cannot handle real valued data directly. It requires discretization and may result in information loss.

2) A.Chouchoulas et.al[4] have discussed about the Quick Reduct algorithm which keeps on adding single attribute at a time and stops when the dependency of the reduced set equals the dependency of original attribute set. The dependency function used by the algorithm is based on positive region of the rough sets. The algorithm was used for reducing the keywords for text categorization. Measure used for selecting the attribute is that of dependency which varies from 0 to 1 and is given as

$$\gamma_P\ (Q) = \frac{POS_P\ (Q)}{|U|} \quad (3)$$

Here, $POS_P\ (Q)$ is the positive region given by rough set theory containing the elements that can be classified in features Q using the information in features P and $|U|$ represents the total number of elements in the set.

Algorithm

Input: C - set of conditional attributes, D - set of decision Attributes

Output: R - Attribute reduct, R ⊆ C

Steps

i) $R \leftarrow \emptyset$
ii) do
iii) $T \leftarrow R$
iv) For each $x\ \in (C - R)$
v) if $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$
vi) $T \leftarrow R\ \cup \{x\}$
vii) $R \leftarrow T$
viii) until $\gamma_R(D) = \gamma_C(D)$
ix) Return R

Here the dependency γ is calculated using eq(3).
Drawback: There is no guarantee of optimal solution as greedy strategy has been used.

3) R.Jensen et.al[5] have discussed about the Fuzzy Rough Quick Reduct algorithm which works in the same way as that of quick reduct algorithm but the dependency function used is based on fuzzy rough sets[5]

4) instead of rough set and is given as

$$\gamma_P\ (Q) = \frac{\sum_{x \in U} \mu_{POS_P(Q)}(x)}{|U|} \quad (4)$$

Here, $\mu_{POS_P(Q)}(x)$ represents the membership of the element in the positive region and it varies from 0 to 1.

Algorithm

Input: $C = \{A_1, A_2 \ldots\ldots A_i \ldots\ldots A_m\}$ is the set of m features of given data set

Output: S is the set of selected features

Steps:
i) $R \leftarrow \emptyset$; $\gamma_{best} = 0$; $\gamma_{prev} = 0$
ii) do
iii) $T \leftarrow R$
iv) $\gamma_{prev} = \gamma_{best}$
v) For each $x\ \in (C - R)$
vi) if $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$
vii) $T \leftarrow R\ \cup \{x\}$
viii) $\gamma_{best} = \gamma_T(D)$
ix) $R \leftarrow T$
x) until γ_{best}=γ_{prev}
xi) Return R

Here the dependency γ is calculated using eq(4).

Drawback: It suffers from the same drawback as that of Quick Reduct algorithm.

5) Qianhua Hu. et.[6]al have discussed about the neighborhood quick reduct algorithm. It is based on the neighborhood model which is used to reduce numerical and categorical features by assigning different thresholds for different kinds of attributes. In this model, the sizes of the neighborhood lower and upper approximations of decisions reflect the discriminating capability of the feature subsets and the size of lower approximation is computed as dependency between decision and condition attributes. A forward feature selection algorithm is used to select the features. It makes use of the measure of dependency which is based on the neighborhood rough sets and is given as

$$\gamma_B(D) = \frac{POS_B(D)}{|U|} \quad (5)$$

Here $POS_B(D)$ is subset of objects whose neighborhood granules consistently belong to one of the decision classes *D* and *B* is a subset of attributes *C* which generates the neighborhood granule.

Algorithm

Input: U- universe, *C*- conditional attributes, *D*-decision attributes, δ - neighborhood controlling parameter

Output: red - reduced Set

Steps

i) $red \leftarrow \emptyset$
ii) for each attribute $A_i \in \{C - red\}$
iii) Compute $\gamma_{red \cup \{A_i\}}(D) = \frac{POS_{B \cup A_i}(D)}{|U|}$
iv) Compute $sig(A_i, red, D) = \gamma_{red \cup \{A_i\}}(D) - \gamma_{red}(D)$
v)End
vi) Select attribute A_k satisfying
$sig(A_i, red, D) = max_i\, sig(A_i, red, D)$
vii) if $sig(A_i, red, D) > \varepsilon$ where ε controls the convergence of sets
viii) $red \leftarrow red \cup A_k$
ix) Goto ii
x) else
xii) Return red
xiii) End if

Here, $sig(A_i, red, D)$ represents the significance of attribute A_i with respect to already selected features and represents the change in dependency when A_i gets added. Dependency γ is calculated using eq(5).

Drawback : Same as that of Quick Reduct algorithm.

6) Hanchuan Peng. et.al.[7] have discussed about the selection of features based on mutual information and making use of the criteria: max-relevance and min-redundancy. Mutual information is defined in terms of probabilistic density. The criterion of max relevance focuses on selecting those features which have maximum mutual information with respect to target class. Thus it tries to maximize the dependency of attribute with respect to target class. But the combinations of individually good features do not lead to good classification performance. Thus to reduce the redundancy, criteria of min redundancy needs to be combined with the criterion of significance. The criterion function used by them is given as:

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j, x) - \frac{1}{m-1} \sum_{x \in S_{m-1}} I(x_j - x_i) \right] \quad (6)$$

Here *I* represent the mutual information between two attributes, S_{m-1} represents the set of selected *m* features and x_j represents the attribute being considered in the current iteration

Algorithm

Input: X - Original Dataset

Output: m-optimal number of features, e^*- classification error

Steps

Step 1: Selecting the candidate feature set

i) Use eq (6) to select *n* sequential features from input *X*. This leads to n sequential feature sets as

$$S_1 \subset S_2 \ldots S_{m-1} \ldots S_m$$

ii) Compare all the feature subsets obtained in step i to find the range of *k* within which the respective error e^k is consistently small.

iii) Find the smallest classification error $e^* = min \quad e_k$ The optimal size of candidate feature set is selected as n^*

Step 2: Selecting Compact Feature Sets

Either forward selection or backward selection can be used for selecting the feature sets. Here we discuss about the forward selection algorithm

i) Set classification error to number of samples

ii) Search for feature subset with 1 feature denoted as Z_1 by selecting the feature x_1^* that leads to largest error reduction.

iii) Then from set{ $S_n - Z_1$ }select feature x_2^* so that the feature set $Z_2 = Z_1, x_2^*$ leads to largest error reduction.

iv) Repeat step iii till the classification error begins to increase.

v) Once the termination condition is reached, selected number of features *m* is chosen as dimension for which lowest error is first reached.

Drawback: The method is based on heuristic approach which does not guarantee global maximization of the criterion function.

3. FEATURE EXTRACTION

This section discusses about the feature extraction algorithms which utilize all the information contained in measurement space to obtain a new transformed space, thereby mapping high dimensional data to lower dimensional one. These algorithms are rich in information but they are time consuming and complex for calculations.

3.1 Principal Component Analysis (PCA)

PCA combines the essence of attributes by creating an alternative, smaller set of variables. The initial data can then be projected onto this smaller set.

The basic procedure of PCA as given in [1] is as follows.

1. The data is collected based on different dimensions.

2. Then the normalization of input data is done so that the attributes fall within range. For this, the mean value of each feature is subtracted from the values of each feature. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains.

3. PCA computes *N* orthonormal vectors which provide a basis for the normalized input data. These are unit vectors that each point in a direction perpendicular to the others. These vectors are referred to as the principal components. The input data is viewed as a linear combination of the principal components. For this it makes use of Covariance matrix, eigenvalues and eigenvectors.

4. Then the use of significance and strength of the principal components is done to sort them in descending order. The principal components essentially serve as a new set of axes for the data, providing important information about variance. That is, the sorted axes are such that the first axis shows the most variance among the data, the second axis shows the next highest variance, and so on. This information helps identify groups or patterns within the data.

5. Since the components are sorted according to decreasing order of significance, the size of the data can be reduced by eliminating the weaker components, i.e., those with low variance. Using the topmost principal components, reconstruction of a good approximation of the original data can be done.

For Example: Fig(2) shows the application of PCA for dimension reduction in weka[10] on the inbuilt glass database.

Drawback of PCA: It is based on the sample covariance which characterizes the scatter of the entire data set, irrespective of class membership. The projection axes chosen by PCA might not provide good discrimination power.

3.2 Linear Discriminant Analysis(LDA)

LDA [8] does more of data classification as opposed to PCA which does more of feature classification. The main goal of LDA is to reduce dimensions by maintaining maximum class discrimination information. This is done by considering two things namely, within-classes scatter and between-class scatter. It works on same lines as that of PCA but makes use of Fischer faces instead of eigen.

Attribute selection output

```
=== Run information ===

Evaluator:     weka.attributeSelection.PrincipalComponents -R 0.95 -A 5
Search:        weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:      Glass
Instances:     214
Attributes:    10
               RI
               Na
               Mg
               Al
               Si
               K
               Ca
               Ba
               Fe
               Type
Evaluation mode:    evaluate on all training data


=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (unsupervised):
        Principal Components Attribute Transformer

Correlation matrix (Step1)
  1     -0.19  -0.12  -0.41  -0.54  -0.29   0.81   0      0.14
 -0.19   1     -0.27   0.16  -0.07  -0.27  -0.28   0.33  -0.24
 -0.12  -0.27   1     -0.48  -0.17   0.01  -0.44  -0.49   0.08
 -0.41   0.16  -0.48   1     -0.01   0.33  -0.26   0.48  -0.07
 -0.54  -0.07  -0.17  -0.01   1     -0.19  -0.21  -0.1   -0.09
 -0.29  -0.27   0.01   0.33  -0.19   1     -0.32  -0.04  -0.01
  0.81  -0.28  -0.44  -0.26  -0.21  -0.32   1     -0.11   0.12
  0      0.33  -0.49   0.48  -0.1   -0.04  -0.11   1     -0.06
  0.14  -0.24   0.08  -0.07  -0.09  -0.01   0.12  -0.06   1

Step 2:Calculate eigenvalues
eigenvalue      proportion      cumulative
  2.51116         0.27902         0.27902       0.545RI+0.492Ca-0.429Al-0.258Na-0.25Ba...
  2.05007         0.22779         0.5068        -0.594Mg+0.485Ba+0.345Ca+0.295Al+0.286RI..
  1.40484         0.15609         0.6629        -0.663K+0.459Si+0.385Na-0.329Al-0.284Fe...
  1.15786         0.12865         0.79155       -0.653Si+0.491Na+0.379Mg-0.276Ca-0.23Fe...
  0.914           0.10156         0.8931        -0.873Fe+0.307K-0.251Ba+0.188Ca-0.154Na...
  0.52764         0.05863         0.95173       -0.657Ba+0.558Na-0.308Mg+0.244K+0.243Fe...

Step 3: Calculate eigenvectors
  V1      V2      V3      V4      V5      V6
  0.5452  0.2857 -0.0869  0.1474  0.0735 -0.1153 RI
 -0.2581  0.2704  0.3849  0.4912 -0.1537  0.5581 Na
  0.1109 -0.5936 -0.0084  0.3788 -0.1235 -0.3082 Mg
 -0.4287  0.2952 -0.3292 -0.1375 -0.0141  0.0189 Al
 -0.2288 -0.1551  0.4587 -0.6525 -0.0085 -0.0861 Si
 -0.2193 -0.154  -0.6626 -0.0385  0.307   0.2436 K
  0.4923  0.3454  0.001  -0.2764  0.1882  0.1487 Ca
 -0.2504  0.4847 -0.0741  0.1332 -0.2513 -0.6572 Ba
  0.1858 -0.062  -0.2845 -0.2305 -0.8733  0.243  Fe

 Step 4: Rank the top attributes and select
  0.721   1 0.545RI+0.492Ca-0.429Al-0.258Na-0.25Ba...
  0.4932  2 -0.594Mg+0.485Ba+0.345Ca+0.295Al+0.286RI...
  0.3371  3 -0.663K+0.459Si+0.385Na-0.329Al-0.284Fe...
  0.2085  4 -0.653Si+0.491Na+0.379Mg-0.276Ca-0.23Fe...
  0.1069  5 -0.873Fe+0.307K-0.251Ba+0.188Ca-0.154Na...
  0.0483  6 -0.657Ba+0.558Na-0.308Mg+0.244K+0.243Fe...

Selected attributes: 1,2,3,4,5,6 : 6
```

Figure 2: Example of PCA for Dimension Reduction

Steps of Working:

i) Consider there are *c* classes. Let μ_i be mean vector of classes for *i=1,...c.* Let M_i be the number of samples within class *i=1,2...c*

$$M = \sum_{i=0}^{c} M_i \quad (7)$$

ii) Two formulae are used by LDA

$$Within\ class\ scatter\ matrix\ (S_w)$$
$$S_w = \sum_{i=1}^{c} \sum_{j=1}^{M_i} (y_j - \mu_i)(y_j - \mu_i)^T \quad (8)$$

$$Between\ class\ scatter\ matrix\ (S_b)$$
$$S_b = \sum_{i=1}^{c} (y_j - \mu_i)\left(y_j - \mu_i\right)^T \quad (9)$$

$where\ \mu = \frac{1}{c} \sum_{i=1}^{c} \mu_i$ is the mean of dataset
μ_i is the mean for class i
y_j is the sample being considered

iii) LDA computes a transformation that maximizes the between class scatter given by eq(9)and minimizes the within class scatter given by eq(8).
iv)The linear transformation is carried out using Fischer Faces
v) The eigen vectors are then computed and top k are selected.
vi)The dataset is then transformed into feature vector.

Drawback: LDA fails to preserve the complex structure of the data which is required for classification.

4. COMBINING FEATURE SELECTION AND EXTRACTION

This section discusses about the algorithm which tries to combine the feature selection and extraction to find a midway between the two basic techniques listed in earlier sections.In [9] P.Maji et.al has proposed a simultaneous feature selection and extraction algorithm which is the first step to combine the merits of both feature selection and feature extraction. It can club the advantages of feature selection techniques which are easy to compute and less time consuming and also the advantages of feature extraction which is that they are rich in content. The algorithmic steps for it are given as

Input : Original set $C = \{ A_1 \ldots . A_k \ldots A_m \}$

Output : Reduced set $S = \{\overline{A_1} \ldots \overline{A_k} \ldots \overline{A_d}\}$

1) Initialise $B \leftarrow \{ A_1 \ldots . A_k \ldots A_m \}\ and\ S \leftarrow \emptyset$

2) Calculate the relevance value $\gamma_{A_i}(D)$ of features $A_i \in B$

3) Select feature A_i from B as the first feature having highest relevance value.

4) Repeat steps (a-d) until $B = \emptyset$ or required number of features are selected

a) Generate 3 subsets with respect to A_i as Insignificant set (I_i) containing least relevant features which can be discarded, significant set (S_i) containing important features and dispensible set (D_i) containing a set of such features which are individually not much relevant but they can be combined to extract new feature which will be rich in information.

b) Extract new feature $\overline{A_t}$ from the features present in D_i and add it to reduced set *S*

c) Discard all the features present in I_i and the used features from D_i from the original set *B*

d) From the remaining features of B, select the feature A_j that maximizes relevance with respect to decision variable as well as maximizes significance with respect to already selected features.

5) End

In step 4, the dispensable set contains the features which may not be individually significant but when combined together to extract a new feature, they would increase the richness of the content. In some iteration the dispensable set may contain only single feature which gets selected. Thus it results in the combination of feature selection and extraction. The relevance and significance are computed using the fuzzy rough theory.

5. CONCLUSION

The paper presents a systematic survey of the dimension reduction techniques. First we discussed about the feature selection strategies and techniques which give a better performance in terms of computational complexity and time. But they have shortfalls that, as some of the dimensions are discarded or not considered; it may result in loss of information and may reduce the accuracy of the classifier. Then we discussed about the feature extraction techniques which give a better performance to have improved classification accuracy but the transformations involved are time consuming and complex. Then we quoted about the simultaneous feature selection and extraction algorithm which is the first step towards finding a midway between the two techniques to combine their advantages. Thus dimension reduction is one the important preprocessing steps which affects the classification/clustering accuracy and also the timing of execution. The different theories like that of rough sets, fuzzy sets, fuzzy rough, mutual information etc are useful for deciding the criteria of rating and choosing the dimensions. Also the problems of missing data, incomplete data are the challenges that must be taken into account during dimension reduction. Hence the combination of selection and extraction techniques, better criterion for evaluating the importance of dimensions and handling the problems associated with real life datasets are the future directions to work upon in this field.

6. REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Technique"s, 2nd ed,The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6.

[2] www.nptel.ac.in

[3] P. Maji and S. Paul, "Rough set based maximum relevance maximum significance criterion and gene selection from microarray data" ,Int. J. Approx. Reason., vol. 52, no.3, pp. 408426, Mar. 2011.

[4] A. Chouchoulas and Q. Shen, "Rough set-aided keyword reduction for text categorization" ,Appl. Artif. Intell.,vol. 15, no. 9, mpp. 843873,Oct.01

[5] R. Jensen and Q. Shen, "Semantics-preserving dimensionality reduction:Rough and fuzzy rough-based approach" ,IEEE Trans. Knowl. Data Eng., vol. 16, no. 12, pp.14571471, Dec. 2004.

[6] Q. Hu, D. Yu, J. Liu, and C.Wu, "Neighborhood rough set based heterogeneous feature subset selection",Inf. Sci., vol. 178, no. 18, pp. 3577-3594, Sep. 2008.

[7] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy",IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226-1238,Aug. 2005

[8] R.O.Duda, P.E.Hart and D.G.Stork, Pattern "Classification and Scene Analysis" ,Hoboken,Wiley,2000.

[9] P.Maji and P.Garai,Fuzzy Rough Simultaneous Attribute Selection and Feature Extraction Algorithm,IEEE Transactions on Cybernetics, VOL. 43, NO. 4,AUGUST 2013

[10] http://www.cs.waikato.ac.nz/ml/weka/