ISSN (e): 2250 – 3005 || Volume, 07 || Issue, 02|| February – 2017 ||
International Journal of Computational Engineering Research (IJCER)

Study of the Class and Structural Changes Caused By Incorporating the Target Class Guided Feature Subsetting in High Dimensional Data

H.N.Meenakshi[1,] P.Nagabushan[2]
[1]Department of Studies in Computer Sceince, Univeristy of Mysore, Mysore, Karnataka, India.
[2]Department of Studies in Computer Sceince, Univeristy of Mysore, Mysore, Karnataka, India.

ABSTRACT
High dimensional data when processed by using various machine learning and pattern recognition techniques, it undergoes several changes. Dimensionality reduction is one such successfully used pre-processing technique to analyze and represent the high dimensional data that causes several structural changes to occur in the data through the process. The high-dimensional data when used to extract just the target class from among several classes that are spatially scattered then the philosophy of the dimensionality reduction is to find an optimal subset of features either from the original space or from the transformed space using the control set of the target class and then project the input space onto this optimal feature subspace. This paper is an exploratory analysis carried out to study the class properties and the structural properties that are affected due to the target class guided feature subsetting in specific. K-nearest neighbors and minimum spanning tree are employed to study the structural properties, and cluster analysis is applied to understand the target class and other class properties. The experimentation is conducted on the target class derived features on the selected bench mark data sets namely IRIS, AVIRIS Indiana Pine and ROSIS Pavia University data set. Experimentation is also extended to data represented in the optimal principal components obtained by transforming the subset of features and results are also compared.

Keywords: *Class of Interest-CoI, target class, feature subsetting, Agglomerative Clustering, Minimum Spanning Tree, K-NN, density, structural property-false acceptance Index, structural property index.*

I. INTRODUCTION

Machine learning and pattern recognition techniques like classification, clustering, regression analysis, and outlier detection are prevailing and further getting optimized in their performance due to the enormous dependency by the various real-world applications. There exist several applications in various domains, which require extracting just the target class from among several classes that are spatially spread at different locations. Consider, for example, mapping the trees in an urban area using the remotely sensed hyperspectral data, to extract the user interested item from an archival that stores several items purchased by the user in an online transaction. Suppose the data provided by these applications are of high dimensional type then it becomes a challenge to the machine learning process thus, necessitate the dimensionality reduction to be used as a preprocessing technique to provide the better visualization of the high dimensional data. In such circumstances, the foremost requirement before automating the applications is to analyze and project the entire high dimensional input data in a feature space that could provide the accurate extraction of the required class called target class. In such circumstances naturally, the control set of the target will be known using which, the most desirable subset of the features could be determined for the better visualization of the target class from among several classes present in the input.

Projection of the entire input space onto a target class guided features subspace is expected to recognize the target class accurately and would also bring certain geometrical or structural changes within the target class as well as other classes present in the input. For instance, if the forest department wishes to observe the area covered by trees from a hyperspectral remotely sensed image, then it is necessary to find the optimal features to accurately map the tree class using its training set. The procedure involved while mapping the tree class would change the other classes like grass, vegetation and water bodies present in the image along with some changes which may have occurred in the tree class too. This paper analyzes the changes in the structural and class properties for the given data so that an appropriate inference can be drawn from the classification results of the target class as well as the other classes after preprocessing. Further, an outcome of the analysis could assist the researcher while solving a multiclass problem by using the target class technique.

Nagabushan et al. [10] have described a procedure to find the optimal subset of the features using the control set of the target class. The work shows that the selection of the low variance features can enhance the compactness within the target class so that its extraction from the input consisting of several classes could be possible. Therefore, employing the method proposed by Nagabushan et al[10] this paper carries out an experimental model to analyze the class and structural properties affected due to the target class guided dimensionality reduction based on the minimum spanning tree (MST), K-nearest Neighbor (KNN) and agglomerative clustering. The minimum spanning tree algorithm is an acyclic subgraph of a graph G is used to address different issues in pattern recognition [1, 2, 3, 4, 5]. In this paper it is used to analyze the structural properties by measuring the changes in the proximity among the samples. In addition to this, the neighborhood changes are also studied using K-Nearest Neighbor (K-NN). The K-NN algorithm is also an extensively used technique for various purposes like classification, clustering[6,7,8,9]. Agglomerative clustering is one of the successfully used clustering approach [10, 11, 12,13]. As an extension to the feature subsetting, principal components are selected from the transformed optimal subset of data and then the analysis is carried out. The analysis carried out in the feature subspace includes the following:

- Structural property analysis through the neighborhood study using the K-NN algorithm.
- Structural property analysis by comparing the distance among the samples in the original feature space with the feature subspace by employing the MST.
- Study on the target class as well as other class property by cluster analysis
- Analysis of the associativity and disassociative among the samples

The remainder of the paper is organized as follows. Section 2 presents the proposed experimentation model and section 3 gives the details of experimentation carried out and is followed by the conclusion.

II. PROPOSED MODEL FOR EXPERIMENTATION:

This section describes the model adopted to study the structural and class properties after incorporating the feature subsetting as decided by the target class. The structural property is described by the geometric structure of the data which depends on the dimensionality of each feature.

Suppose $\mathbf{T_C} = \left\{ \mathbf{s_1}, \mathbf{s_2}, \dots, \mathbf{s}_t \,\middle|\, s_i \in \Re^N \right\}$ is the control set of the target class then, by employing the feature subsetting proposed by Nagabushan et al.[10] all the features with large variance that would tend to split the target class will be eliminated and from the remaining features optimal features are obtained which could project the target class accurately by maximizing the compactness of the given class so that $\mathbf{T_C} = \left\{ \mathbf{s_1}, \mathbf{s_2}, \dots, \mathbf{s}_t \,\middle|\, s_i \in \Re^n \,\middle|\, n << N \right\}$. Given $X = \left\{ x_1, x_2, \dots, x_M \,\middle|\, x_i \in \Re^N \right\}$ the input from which the target class requires to be extracted, and then $\{ X \}$ also get projected on to the n optimal sub set of features.

Since the transformation of the features could further enhance the projection of the data, in continuation to the subsetting optimal principal components can be determined by transforming the feature sub space. Suppose the diagonalization of a covariance matrix of target class training set results in n Eigen values that are represented in an increasing order as $\lambda = \{ \lambda_1, \lambda_2, \dots \lambda_n \}$ and $\{ EV \}$ is the set of corresponding eigen vectors. In conventional linear transformation technique like PCA, the data space can be projected onto orthogonal principal components that preserve 90%-95% of variance [14]. In a similar manner to conventional PCA if the higher order principal components are selected for the projection of the CoI then it would tend to split the CoI itself. Thus, the principal components are computed from the eigen values that does not tend to spilt the target class. Therefore, if the largest eigen values whose variance is greater than 60%-65% is discarded and the next eigen values are selected until the density within the target class remains same. As a result, an optimal subset of eigenvalues and its corresponding eigenvectors between the range $\left[\underline{\lambda} - \overline{\lambda} \right]$ form the optimal eigen subspace and are most significant in computing the principal component. $\overline{\lambda}$ is the upper bound of the eigenvalue(largest eigenvalue) that is not greater than 60%-65%. $\underline{\lambda}$ is the lower bound of the eigenvalue that results in the maximum homogeneity within the target class beyond which the homogeneity doesn't change and results in

$\mathbf{T_C} = \{ \mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_t \mid s_i \in \Re^d \mid d << n \}$ According to the availability of the input, it can be transformed by using the principal components derived from the target class eigen subspace as shown in equation 2.

$$X' = [X]\left([EV]^T\right)^{-1} \qquad \text{eq (1)}$$

It is evident from the equation (1) that the transforming of the input based on the eigenvectors obtained from the target class eigen subspace would definitely help in the categorization of the target class samples from the input space. Nevertheless, the intriguing points of the study are the changes in the neighborhood property of the input space and the class property of the classes present in the input.

2.1. Analysis of structural properties

The training phase required to learn about the target class for the purpose of feature selection is based on just the target class samples. In the original feature space, the initial intuition is that all the neighbor samples of every training sample belong to the same class. However, this neighborhood might not continue to exist for the reason that the input samples of any other class would become the neighbor of any sample belonging to the target class that requires being verified. Therefore, the experimentation model to analyze the structural property should validate the following issues:

1. The resultant total distance calculated by connecting all the training samples based on their shortest proximity in the intrinsic feature subspace obtained during the training phase should be less than or equal to the distance in the original feature space.
2. Classification of the input space by incorporating the feature reduction either by subsetting or transformation guided by the target class training set should ensure that the neighborhood of target class samples will have only the samples that belong to the target class but not any other samples belonging to any unknown classes.

The first issue focuses on the quality of the training phase and the second phase addresses the classification results. Unless there is a favorable geometric structure change in the target class represented in the optimal feature subspace, the classification accuracy of the target class cannot be guaranteed. The geometric structural changes can be treated as favorable only when the subspace results in a good classification results in terms of minimum false acceptance and maximum true acceptance into the target class. In view of the geometric of data K-Nearest Neighbor is estimated to analyze the neighborhood changes[ ] and a graph based Minimum Spanning Tree (MST) approach is also employed to evaluate the total distance between the samples of the given class before and after its feature reduction.

2.1.1. K-Nearest Neighbor (KNN): To represent the K-nearest neighbors of all M training samples the NN-matrix have to be computed in which samples are ordered row-wise with each column consists of k^{th} NN. For M samples and K neighbors the NN matrix is MxK. The following acronym is used with respect dimensionality reduction of the data:
OFS- Original feature space, FSS- Feature Subset Space and TFS- Transformed Feature Subspace.

K-NN can be employed to analyze the impact of optimal features on the samples at two phases:

Phase 1: The structural changes caused within the target class are measured in terms of the neighborhood changes during the training phase in the OS, *FS*, and *TS*.

Phase 2: Analysis of neighborhood changes in target class after classification of input space.
The structural property that exists in the original feature space may not remain the same after feature reduction in the training phase and after classification thereafter. In the first phase, the overall percentage of neighborhood properties being preserved for all of the K-NN is measured. In the second phase, the order of the K-NN samples gets preserved for the target class in the reduced space after classification to ensure true acceptance.

Phase1: To measure the amount of structural property being preserved, an indicator called structural property index is here proposed which is determined as follows. If the k^{th} neighbor of the sample S_i has changed its position to either far or near after feature reduction then each row of NN matrix in OFS can be compared with the corresponding row of the NN matrix obtained in the reduced feature space. Results of such a comparison indicate that the change or transformation caused in the neighborhood can be represented in the form of a new matrix called Transformed NN matrix.–TNN matrix. Each row of the TNN matrix reflects the shift of the neighborhood of a sample S_i to either far or near after feature reduction. The procedure to compute the change in position is as explained in procedure given below. If the new position of any kth neighbor of a sample S_i is a positive number it then indicates that the neighbor of that sample has moved far away, a negative number indicates that sample has changed to a nearby position and zero indicates that no changes have occurred due to feature reduction.

Algorithm to compute the change in position

Input : $\{Tnn\}_{MxN}$, $\{Tnn'\}_{Mxn}$ NN matrix of target class computed in original feature and in reduced feature sub space

Output: TNN matrix consisting of weight indicating the structural change

for i=1 to M **do** // for M samples
  for x =1 to K **do** // for K nearest neighbors of samples S_i
    P= x neighbor of S_i in an original feature Space;
    for y =1 to K **do**
      Search for P in NN matrix of S_i in feature subset space;
      if found $P'_{ix} = y - x$;
    end;
  end;
end;

Depending on the changes represented in a new matrix TNN, the ***structural property index*** is computed by assigning credence to the change. Zero change is given with highest credence, and the next change is with less credence. If W_i is the credence for zero change then $\{\pm 1\}$ change gets W_{i-j} . For example, if zero change means 100% credence then $\{\pm 1\}$ change means 75%. Similarly, the credence is assigned to cover all KNN being computed. Then the structural property index is given as in (2).

$$StructuralIndex = \mathrm{I} = \frac{\sum_{i=1}^{K} W_i}{K} \qquad \text{eq (2)}$$

The index varies between [0 1], where I=0 means 100% structural property is preserved or in the other way no structural changes has caused due to feature reduction.

Phase2: Let $\{ S_1, \ldots S_j \mid j = 1 \ldots K \}$ is the K-NN of a sample S_i in the original feature space. After the feature reduction, if the same order of the neighborhood is maintained for the sample S_i then, it suggests that no structural changes have happened to the target class during the training phase. After classifying the input samples as target class if K nearest neighbors of the sample S_i further continue to exist then the feature reduction has not caused the structural property to change and also has not allowed a false acceptance into the target class.

Suppose the order of the neighborhood has changed and the new neighbors of the samples S_i are also found to be belonging to the target class then such structural changes caused by feature reduction are favorable and acceptable. On the other hand, if the new neighbors of a sample S_i are from some unknown classes then it follows that the intrinsic feature subspace obtained from the target class dimensionality reduction have destructed the geometric structure of the target class but induced a false acceptance into the target class which is not acceptable. The structural changes along with false acceptance can be measured by an index called ***structural property-false acceptance Index-*** H **.** If is H zero then there may be or may not be structural changes but there is no false acceptance into the target class which is as expected by the application.

Algorithm to compute the structural property-false acceptance Index:

Input: $\{Tnn'\}_{Mxn}$, $\{Inn'\}_{Xxn}$: K-NN matrices of target in reduced feature space and input space represented in the target class derived feature space including the target class training set

Output:

for i=1 to M **do** // for M samples

$H = 0$; // structural property-false acceptance Index

for j =1 to K **do** // for K nearest neighbors of samples S_i

if (j^{th} nearest neighbor of sample S_i in $\{Tnn'\}$ $\neq$ j^{th} nearest neighbor of sample S_i in $\{Inn'\}$) then,

if (j^{th} nearest neighbor of sample S_i in $\{Inn'\}$ $\notin$ $\{Tnn'\}$ then,

$H++$;

end;

end;

Another study on the structural property using the K-NN is to find out the number of K common nearest neighbors for a sample represented in one feature space when compared to another feature. By varying the value of K the neighborhood changes can be measured. Suppose $f(s_i) \rightarrow \{nn_1, \dots nn_K\}$ is a function applied on a sample to find the K-NN in a given features space then the common neighbors can be analyzed $g(f(s_i), f(s_i'))$ where g is function to find the common neighbors of a samples in two different feature space.

2.1.2. Minimum spanning Tree (MST)

MST is constructed to validate the changes caused in the proximity among the samples due to dimensionality reduction. Given an undirected weighted graph $G = (V, E, w)$ with non-negative weights, the minimum spanning tree (MST) problem is to connect all the nodes that have the minimum weight. Each sample from the target class training set is considered as a vertex in N dimensional and the weight is computed between the vertices as in equation (3). Similarly, pairwise distance between all the samples can be computed to form the graph G which represents the target class.

$$d_i(x_a, x_b) = \sqrt{(x_a - x_b)^2} \qquad \text{eq (3)}$$

Further, the Prims algorithm is used to construct the MST in N-dimensional space. Similarly, after the feature reduction graph is built in both *L*-dimensional and *n*-dimensional space followed by MST in the intrinsic subspace. To analyze the changes in proximity two MSTs are required to be compared. Let T_i and T_j are the MST obtained in *L* and *n* dimension respectively. The distance between T_i and T_j is the edges present in $E(T_i) \oplus E(T_j)$ (symmetric difference) given as

$$d(T_i, T_j) = \left| E(T_i) \oplus E(T_j) \right| \qquad \text{eq (4)}$$

A greater difference indicates several changes in the neighborhood and changes in the proximity between the samples whereas a small difference indicate the symmetric or isomorphism of the two MST.

2.2. Analysis of Class properties

The optimal features obtained from feature subsetting or principle components derived from target class eigen subspace are used to reduce the input space. When such target class guided dimensionality reduction of input is subjected to classification, it is expected that the samples belonging to the target class would get classified accurately. In this context, it is essential to study the impact of feature reduction on the samples of unknown classes that are present in the input which could affect the classification results of the target class. To simplify the analysis, a training set of other classes is also assumed and an agglomerative clustering is adopted to analyze the class behavior of each sample due to dimensionality reduction. The analysis of class property based on the clustering is carried out at three levels:

Level1: Clustering of the target class training samples in *N*-dimensional original feature space, *L*-subset of features.
Level2: The input projected onto the target class derived feature subspace when subjected for classification that results in several classes. The input samples that are classified as target class in the reduced feature space which are carried for cluster analysis.
Level3: The other classes resulted in the classification also undergoes cluster analysis.
The same principle of the clustering algorithm is adopted in all the three level. It starts with each sample of a class as a cluster. Clusters are further combined based on their close proximity to the samples. The proximity is computed based on Euclidean distance as it allows the summarizing of the collection of samples by their centroid which is the average of the samples. The merging continues till all the samples are grouped into one cluster. Once all the samples are grouped into one cluster in the form of a tree, the inconsistent samples of the cluster can get estimated. Inconsistent samples are measured to determine the samples that are more deviated from the centroids of the cluster. Such deviated samples are the indication of samples or sub-clusters that get delayed in getting merged due to more proximity or dissimilarity. The cluster tree is cut at various random points to find the inconsistent samples by using the equation (5). The inconsistent value of k+1 cluster is a measure of the separation between the two clusters whose merger is represented by that node, compared to the separation between sub-clusters merged within those clusters which are normalized by the standard deviation of that cluster.

$$I_{k+1} = (d(l^{i}_{k-1}, l^{j}_{k-1}) - \mu_k) / \sigma_k \qquad \text{eq (5)}$$

The feature reduction process is expected to minimize the inconsistent samples within the target class. Therefore, in level1 cluster analysis has to be carried out by considering the training set of target class before and after feature reduction so that standard deviation can be minimized. In level2 the input samples that are classified as target class are considered for cluster analysis so that true acceptance as well as false acceptance into the required class can be validated. Similarly, in level3 samples classified under other classes are subjected to cluster analysis so as to verify the impact of target class derived features subspace. Since the samples of other classes are represented in the target class feature space it is expected that they don't get classified into the target class. Suppose the input contains any overlapping samples, then the target class feature subspace would become a closely relevant feature space of overlapping classes. Therefore, if inconsistent samples result when the cluster tree is cut off, it then indicates that dimensionality reduction is effective in maximizing the distance between target class and the other classes.

III. EXPERIMENTS AND RESULTS

The experimental analysis was carried out on four bench mark data sets as described in table1. The experimentation was conducted individually on different target classes chosen from the data set listed in table1.

Table1. Data set description used for experimental analysis of structural and class property

Data set	Dimension	classes
IRIS data set	4	3
AVIRIS Indian Pine mini data set	200	8
ROSIS Pavia University mini data set	103	5

3.1. Iris data set

Setosa as CoI: Considering Setosa as the target class both feature subsetting and eigen subspace was employed. To find the subset of optimal feature original setosa class 10 training samples were considered initially as listed in table 2a. The features whose variance was greater than the threshold value were eliminated. The threshold was chosen to be slightly greater than the smallest variance that would not split the target class. 2 features having variance greater the threshold were eliminated and features 3, 4 were chosen as the optimal subset. Similarly, the eigen subspace was found from the original space of the seotsa class and two major principle components were selected to represent the target class. Subsequent to finding the optimal subset of features and principle components structural, the class properties were analyzed.

3.1.1. Structural Analysis

KNN: Assuming K=5, the nearest neighbor matrix- KNN Matrix is computed in the original feature space, feature subset space and eigen subspace of setosa class as in table 2b. The sequenced number of the samples according to their occurrence in the input space is used as entries in the matrices. For the purpose of illustration, the first 10 samples are being considered.

Table 2a. 10 training samples of Setosa class in original feature space

Sample	F1	F2	F3	F4
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.7	0.2
5	5.4	3.9	1.7	0.4
6	5.1	3.5	1.4	0.2
7	4.9	3.1	1.5	0.1
8	4.4	2.9	1.4	0.2
9	5.0	3.4	1.5	0.2

Based on the NN matrix, the geometric structure changes were observed by comparing the NN matrix of original features space-OS against the NN matrix obtained after feature subsetting and the results of the comparison were noted in the form of the transformed NN matrix-TNN as in table 2c. It was observed from the TNN matrix that some of the samples were moved forward that is indicated by a positive number and that some of the samples had become near neighbor as indicated by the negative number. Values '0' indicates that a few samples had not changed their neighborhood. Some samples were found to be not within 5NN which are indicated by ∞ . This signifies that even after representing the Setosa class with its optimal features, the geometric structure of some of the samples has not changed. From the table 2c, the *structural property index* is computed so as to determine the average structural property being preserved by each sample. The computation of the index is based on the weight assigned to every change in the neighborhood. The weight is assumed as follows:

Change in neighborhood position	weight [1-0]
'0'	1
{+1} or {-1}	0.8
{+2} or {-2}	0.6
{+3} or {-3}	0.4
{+4} or {-4}	0.2
∞	0

Using equation (3) the average index computation is given below

S	Structural property index in feature subset space	Structural property index in eigen subspace
1	(0.8+0+0+1+1)/5= **0.56**	(0.6+1+0.6+0.4)/5=**0.52**
2	(0+0.6+0+0+0.6)/5= **0.24**	(1+0.6+0.6)/5=**0.44**
3	(0+0.4+0.8+0+0.8)/5= **0.4**	(0.2)/5=**0.4**
4	(0+0+0.8+0+0.8)/5= **0.32**	(0.4+1)/5=**0.28**
5	(1+0+0.6+1+0.8)/5= **0.52**	(0.8+0.6+0.6+0.4)/5=**0.48**
6	(0+0.4+0.8+0.8 +0.8)/5= **0.56**	(1+0.8)/5=**0.36**
7	(0.2+0+0+0.8+0)/5= **0.2**	(0.6+0.6)/5=**0.24**
8	(0.6+0.4+0.8+0+0.8)/5= **0.52**	(1+0.6)/5=**0.32**
9	(0+0.6+0.8+0.8+0)/5= **0.44**	(1+0.6)/5=**0.32**
10	(0.8+0+0.6+0.6+0.6)/5= 0.44	(1+0.6+0.6)/5=**0.44**

In an average, the geometric structure maintained by the 10 training samples after the feature subsetting and transformed eigen subspace is **42%** and **38%.** The favorability of the feature subsetting can further be verified with the results of the classification. Next, the entire IRIS data set with 150 input samples (which also includes the training samples that are considered while training about the setosa class) were projected on to the features derived from setosa target class and after that, an NN matrix is computed.

Table 2b. NN matrix in original space, Feature Subset space, in eigen subspace

S	1-NN	2-NN	3-NN	4-NN	5-NN	S	1-NN	2-NN	3-NN	4-NN	5-NN	S	1-NN	2-NN	3-NN	4-NN	5-NN
1	5	8	10	3	7	1	2	5	9	3	7	1	3	8	5	2	10
2	10	3	4	8	9	2	1	5	9	3	7	2	10	8	1	6	4
3	4	7	2	10	8	3	1	2	5	9	7	3	5	1	8	9	4
4	3	9	10	7	2	4	8	10	1	2	5	4	6	9	8	3	5
5	1	8	7	3	10	5	1	2	9	3	7	5	3	1	9	8	7
6	5	1	8	7	10	6	4	8	7	10	1	6	4	9	8	10	3
7	3	4	8	5	10	7	1	2	5	9	3	7	9	5	3	4	6
8	1	5	10	3	2	8	4	10	1	2	5	8	1	9	4	6	10
9	4	3	2	7	10	9	1	2	5	3	7	9	4	6	8	3	5
10	2	3	4	8	1	10	4	8	1	2	5	10	2	8	1	6	4

Table 2c. Comparison of NN matrix of original space with feature subset space and with eigen subspace in TNN matrix

S	1-NN	2-NN	3-NN	4-NN	5-NN	*S*	1-NN	2-NN	3-NN	4-NN	5-NN
1	+1	∞	∞	0	0	*1*	2	0	2	-3	∞
2	∞	+2	∞	∞	-2	*2*	0	∞	2	-2	∞
3	∞	3	-1	∞	-1	*3*	4	∞	∞	∞	-2
4	∞	∞	-1	∞	-1	*4*	3	0	∞	∞	∞
5	0	∞	+2	0	∞	*5*	1	2	2	-3	∞
6	∞	+3	-1	-1	-1	*6*	∞	∞	0	∞	-1
7	+4	∞	∞	1	∞	7	2	2	∞	-2	∞
8	2	3	-1	∞	-1	8	0	∞	2	∞	∞
9	∞	2	-1	1	∞	9	0	2	∞	∞	∞
10	3	∞	-2	-2	-2	10	0	∞	2	-2	-2

Similarly, the input was transformed to the eigen subspace derived by the Setosa class and then NN was obtained as is given in table 2d. Further, the nearest neighbors computed for 10 setosa training samples in the feature subset space during the training phase given in table 2b is compared with the nearest neighbors obtained for the same 10 samples after the classification as given in table 2d. This comparison helps us to understand the neighborhood changes caused due to misclassified samples of the target class.

Table 2d. Comparison of NN after classifying the IRIS 150 input samples based on the feature space obtained from Setosa as target class

Feature subset space						eigen subspace					
S	1-NN	2-NN	3-NN	4-NN	5-NN	*S*	1-NN	2-NN	3-NN	4-NN	5-NN
1	22	10	9	3	5	*1*	2	5	10	3	12
2	1	5	18	34	25	*2*	1	5	21	3	7
3	1	6	5	8	14	*3*	8	21	5	9	7
4	21	6	1	2	5	*4*	9	10	11	42	49
5	35	23	17	12	32	*5*	1	36	22	17	27
6	14	8	7	10	1	*6*	4	13	45	10	34
7	12	13	5	4	27	7	7	2	5	9	3
8	45	29	8	2	5	8	14	10	1	2	5
9	9	21	5	3	41	9	3	2	5	3	7
10	16	11	9	2	5	10	7	8	9	2	43

Similarly, the samples in the transformed eigen subspace were also compared for nearest neighbors during training phase and after classification. The comparisons were noted in the form of structural property-false acceptance index which is computed as follows:

S	$H_{\text{In feature subspace}}$	$H_{\text{In eigen subspace}}$
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0
9	0	0
10	0	0

The value of H for 10 samples of setosa class was found to be zero in both of the feature reduced spaces. This indicated that, there were no false acceptances into the target class irrespective of structural changes.

MST: Similar to KNN, Minimum Spanning Tree was also constructed on the original feature space, the optimal feature subset and the eigen subspace derived principle components for 10 of the training samples of the setosa class. To build a tree, a pairwise Euclidian distance was computed for all of the 10 training samples as in (3). Next, the distance between the MST obtained in OS-FS, OS-TS were computed. Later, considering the complete 150 IRIS data set the MST was built. The input was projected once onto the setosa derived optimal feature subset and separately onto the eigen subspace. For each case, the MST was built separately. The distance between MST of all 150 samples in OS-FS and OS-TS were then calculated for comparison. A similar experiment was carried out by varying the training samples from 10 to 50 samples and for individual training set the value of K was varied for KNN analysis and MST was also build for distance comparison. The table outlines the structural analysis results for the different training set.

Versicolor as CoI: Since, versicolor has many overlapping samples with verginica; a structural analysis by means of KNN and MST was carried out by considering the 10 training samples of versicolor as listed in table 3a.

KNN: As with the KNN procedure as demonstrated above for setosa class, the following procedure lists the result for versicolor class treated as target class. Feature 1 and feature2 get eliminated so that only feature 3 and feature 4 were being considered to be the optimal subset for versiciolor class. Similarly, versicolor in the original feature space was transformed, which resulted in 2 principle components. Table 3b shows the NN matrix computed for versicolor.

Table 3a. 10 training samples of Setosa class in original feature space

Sample	F1	F2	F3	F4
51	7.0	3.2	4.7	1.4
52	6.4	3.2	4.5	1.5
53	6.9	3.1	4.9	1.5
54	5.5	2.3	4.0	1.3
55	6.5	2.8	4.6	1.5
56	5.7	2.8	4.5	1.3
57	6.3	3.3	4.7	1.6
58	4.9	2.4	3.3	1.0
59	6.6	2.9	4.6	1.3
60	5.2	2.7	3.9	1.4

Table 3b. NN matrix computed in original space, Feature Subset space, in eigen subspace for target class versicolor

S	1-NN	2-NN	3-NN	4-NN	5-NN	S	1-NN	2-NN	3-NN	4-NN	5-NN	S	1-NN	2-NN	3-NN	4-NN	5-NN
51	53	59	52	55	57	51	59	55	57	56	52	51	58	59	53	52	55
52	57	59	55	51	53	52	55	56	59	51	57	52	60	57	51	58	53
53	51	59	55	52	57	53	51	57	55	59	52	53	55	59	54	51	56
54	60	56	58	55	52	54	60	56	52	59	55	54	55	53	59	56	51
55	59	52	57	53	51	55	52	51	57	59	56	55	53	54	59	51	56
56	54	60	55	52	57	56	59	52	55	51	57	56	60	53	57	55	52
57	52	55	59	53	51	57	55	51	52	53	59	57	60	52	56	53	51
58	60	54	56	52	55	58	60	54	56	52	59	58	51	59	52	53	55
59	55	52	53	51	57	59	56	51	55	52	57	59	55	53	51	54	58
60	54	56	58	52	55	60	54	56	52	59	55	60	57	52	60	53	55

Table 3c. Comparison of NN matrix of original space with feature subset space and eigen subspace in TNN matrix form

S	1-NN	2-NN	3-NN	4-NN	5-NN	*S*	1-NN	2-NN	3-NN	4-NN	5-NN
51	∞	-1	2	-3	-2	51	2	0	1	1	∞
52	4	1	-2	0	∞	52	1	∞	∞	-1	0
53	0	2	0	1	-3	53	3	0	-2	∞	∞
54	0	0	∞	1	-2	54	∞	2	∞	-3	∞
55	3	-1	0	∞	-3	55	2	∞	∞	-3	-1
56	∞	∞	0	-2	0	56	∞	-1	1	1	-2
57	2	-1	2	0	-3	57	1	∞	∞	0	0
58	0	0	0	0	∞	58	∞	∞	∞	-1	0
59	2	2	∞	-2	0	59	0	∞	-1	-1	∞
60	0	0	∞	-1	0	60	∞	∞	∞	-2	0

Table 3d. Comparison of NN after classifying the IRIS 150 input samples based on the feature space obtained from Setosa

Feature subset space						eigen subspace					
S	1-NN	2-NN	3-NN	4-NN	5-NN	S	1-NN	2-NN	3-NN	4-NN	5-NN
51	60	57	61	56	101	51	68	59	53	67	105
52	65	56	59	51	150	52	62	57	59	81	120
53	51	75	55	73	150	53	55	59	54	51	152
54	60	67	69	90	134	54	55	53	99	88	89
55	52	51	57	59	120	55	93	54	64	67	56
56	92	96	89	51	76	56	72	53	57	55	129
57	55	51	52	53	100	57	80	52	56	53	51
58	66	54	56	52	59	58	61	59	52	53	67
59	56	78	55	52	57	59	65	53	51	61	81
60	59	71	63	81	105	60	78	89	100	92	74

Further, the NN matrix in the original feature space was compared to the NN matrix obtained in the feature subset space and the NN matrix in the original feature space with eigen subspace principle components to study the changes in the neighborhood. Table 3c gives the results of the comparison. Based on the NN matrix as given in table 3c, the structural property index- I was computed as given below:

S	Structural property index in feature subset space	Structural property index in eigen subspace
51	(0+0.8+0.6+0.4+0.6)/5= **0.48**	(0.6+1+0.8+0.8+0)/5=**0.64**
52	(0.2+0.8+0.6+1+0)/5= **0.52**	(0.8+0+0+0.8+1)/5=**0.52**
53	(1+0.6+1+0.8+0.4)/5= **0.76**	(0.4+1+0.6+0+0)/5=**0.4**
54	(1+1+0+0.8+0.6)/5= **0.68**	(0+0.6+0+0.4+0)/5=**0.2**
55	(0.4+0.8+1+0+0.4)/5= **0.52**	(0.6+0+0+0.6+0.8)/5=**0.4**
56	(0+0+1+0.6 +1)/5= **0.52**	(0+0.8+0.8+0.8+0.6)/5=**0.6**
57	(0.6+0.8+0.6+1+0.4)/5= **0.68**	(0.8+0+0+1+1)/5=**0.56**
58	(1+1+1+1+0)/5= **0.80**	(0+0+0+0.8+1)/5=**0.36**
59	(0.6+0.6+0+0.6+1)/5= **0.56**	(1+0+0.8+0.8+0)/5=**0.52**
60	(1+1+0+0.8+1)/5= **0.76**	(0+0+0+0.6+1)/5=**0.32**

S	H In feature subspace	H In PC
51	1	1
52	1	1
53	1	1
54	1	0
55	1	0
56	0	1
57	0	0
58	0	0
59	0	0
60	1	0

In the next step, all 150 samples of the IRIS data set were represented on the feature space derived by the versicolor target class and then classified to map the versicolor class.

An NN matrix was computed on the samples classified as versicolor as shown on table 3d. Further, the structural property-false acceptance Index H was also computed to study the falsely accepted samples after feature reduction. The value of H =1 indicates that for the sample S_i one of the nearest neighbor does belong to an unknown class hence, it is a false acceptance into the target class with or without structural changes. A false acceptance could exists at K>5. But in this illustration K values is chosen to be 50% of the training size.

The nearest neighbor analysis was also used to study the common neighbors by varying the K value in different feature space. Also, the % of the sequence of the order of the neighborhood maintenance is computed which is shown in fig 1.

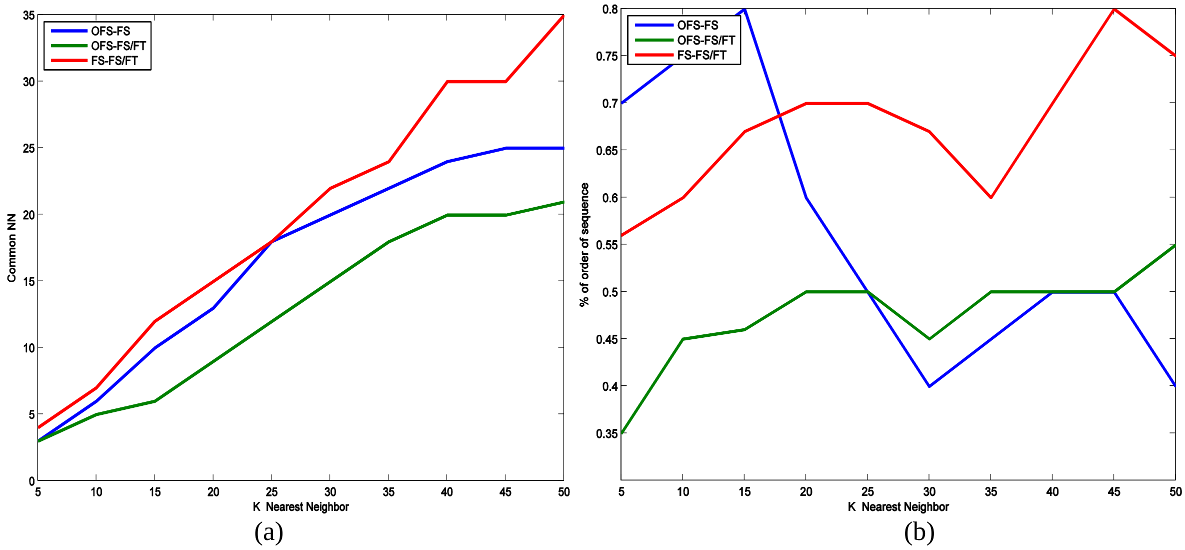


Fig1. (a) K-Common neighbors of a sample (b) % of order of sequence maintained for IRIS data set

MST: A minimum spanning tree was built on the original, feature subset and eigen subspace of the versicolor training samples. The distance between these MST was then computed. After classification the MST was also built on the reduced space for target class and the results are as given in table 3e.

3.1.2. Analysis of Class properties

Agglomerative clustering was used on training samples of Setosa class and inconsistent samples were measured as in (5) to see how many samples deviated from the mean in original space. Further, the same samples were represented in the optimal subset of feature space and next on principle components. In each intrinsic subspace, clustering was carried out to see inconsistent samples. It was observed that the standard deviation of all the samples was minimized in subspace when compared to the original space. Further, the clustering was done on the entire input space consisting of 150 samples projecting on to setosa feature space. Since, this experimentation was set up for the validation purpose; the training set of other two classes namely, Verginica and Versicolor was used to classify the entire 150 samples in the Setosa class derived feature subspace. On the classified class, cluster analysis was implemented so as to realize the behavior of verignica and versicolor samples. It was observed that, by cutting the cluster tree of setosa class randomly at different heights, the tendency of the split of the samples was less whereas for other two classes there were more inconsistent samples.

Table 3e. Structural analysis of IRIS data set based on KNN and MST

Target class	Training samples	KNN(K=50%of training set)				MST			
		Feature subsetting		Eigen subspace		Training phase		After classification	
		I	H	I	H	d(OFS,FSS)	d(OFS,TFS)	d(OFS,FSS)	d(OFS,TFS)
Setosa	10	42%	0	38%	0	4	4	36	37
	20	46%	0	41%	0	9	10	36	36
	30	50%	0	40%	0	12	15	38	38
	40	48%	0	40%	0	25	30	32	32
	50	51%	0	41%	0	30	38	34	34
Versicolor	10	16.52%	12	14.21%	10	2	4	12	12
	20	13.21%	10	10.12%	9	5	7	14	14
	30	11.34%	8	9.36%	7	9	10	13	13
	40	9.99%	8	8.67%	8	12	12	14	14
	50	8.0%	9	6.23%	10	13	13	12	12

3.2. Analysis of the high dimensional data in the reduced feature space for class and structural properties

Since the proposed dimensionality reduction is focused towards high dimensional data; a similar analysis was carried out on Pavia University and Indiana Pine data set which has high spectral bands. Since Pavia University and Indiana pine data set consists of more number of samples and classes, a mini subset was created for experimental analysis whose details are as given in table 4. There were a lot of correlated spectral bands in both the data sets and hence, feature elimination found good scope in eliminating correlated and high variance spectral bands. The number of features that got eliminated from the original features space followed by the optimal subset of features being selected and principal components obtained from the eigen subspace are listed in the table5a. The results reported in the table are of an experiment conducted by assuming each class as a target class.

Table 4. Mini- Indiana Pines data set Total 500 samples

	AVIRIS Mini- Indiana Pine data set				ROSIS Pavia University data set			
	Class	Samples in original data set	% of samples in original data set	No of samples for 500	Class	Samples in original data set	% of samples in original data set	No of samples for 300
1	Alfalfa	46	0.45%	3.530315	Asphalt	6631	19.09%	57
2	Corn-notill	1428	14%	109.5932	Meadows	18649	53.71%	162
3	Grass-trees	730	7.12%	56.02456	Trees	3064	8.825%	26
4	Stone-Steel-Towers	93	0.9%	7.137375	Painted metal sheets	1345	3.87%	12
5	Oats	20	0.195%	1.534919	Bare Soil	5029	14.48%	43

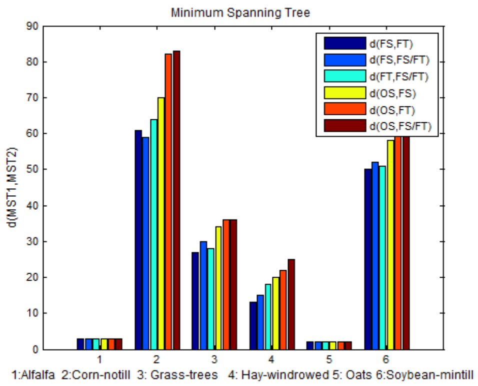
On the reduced dimensional space area of the entire input space H was also measure when subjecting for classification the. The structural property index was observed to be more than 50% and H was seen to be very less. This indicates that there was a structural change in the target class during the training phase as well as after the classification. However, the false acceptance into the target class was very negligible. It can be inferred that although the target class dimensionality reduction has caused a geometric change in the structure it has not yet destroyed the classification results. The proposed dimensionality reduction techniques were tested for the classification of the target class. Classification results of the entire input space after its projection on to the optimal features and eigen subspace as decided by the chosen target class is as shown in table 6. Sensitivity or recall- ρ is used to test the number of samples that are truly classified as the target class and the number of samples classified as target class that is relevant using precision- γ as given in (7).

$$\rho = \frac{TP}{(TP + FN)} \text{ and } \gamma = \frac{TP}{(TP + FP)} \qquad \text{eq (7)}$$

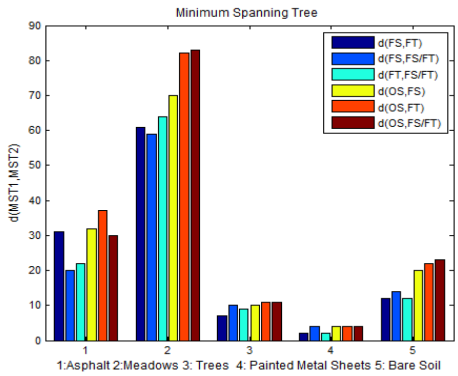
Further, the minimum spanning tree was also built on the various target classes chosen from both the high dimensional mini data set. The distances, measured between each MST in different feature space are shown in fig2. The distance between the spanning trees constructed in the original space is compared with the tree built in the feature subset space as well as the tree in the transformed eigen subspace. The distance between the spanning tree in the OFS, FSS, and OFS, TFS was comparatively higher which indicates that the changes proximity between the samples in the original feature space is not the same after incorporating the feature reduction process.

KNN was also employed to analyze the structural property of hyperspectral data. Structural property index was computed to analyze the structural changes. The values of **I** for none of the chosen target class was found to be zero which indicates that the both the feature subsetting and feature transformation has caused structural changes in the data. To understand the favorability the classification results were also taken into consideration. The value of **H** indicated that the false acceptance was found along with structural changes. Particularly, for the target class Oats, the value of was **H** large which is greater than the actual number of samples. One of the reasons is due to the very less number of training samples chosen in the mini data set. However, the study of the optimal number of training samples for the dimensionality reduction is out of the scope of this paper which can be considered as separate future work.

In continuation of the structural analysis, cluster analysis was carried out to study the class property after the classification of the input space based on the chosen target class feature space. Fig 3 illustrates the dendrogram plotted for the target class Meadows from the Pavia university mini data set but the results of the other classes are not shown in the paper. The cluster tree was cut at different levels to analyze the inconsistent samples. The cutoff point- C was chosen as being less than the standard deviation of the class 0<C<1 and then inconsistent samples were measured as in (6). The experiment was conducted for several cutoff levels. The cluster tree was cut at different levels until all the nodes or samples are split to form a single sample. It was observed that some of the samples that were delayed in the merging with the tree in the original feature space did not tend to get split in the reduced feature space. This indicates that the optimal features obtained in the target class guided dimensionality reduction process have improved the binding between the samples belonging to the same class.



2a



2b

Fig 2. Structural Property Analysis of target class using MST: 2a. Indiana Pine mini dataset 2b. Pavia University mini data set:

Table 5. Structural analysis of high dimensional data after the target class guided dimensionality reduction

Data Set	Target class	Structural analysis after Feature Subsetting				Structural analysis in Eigen Subspace		
		Feature subsetting		KNN with 30% of training samples for K=50%		PC's	KNN with 30% of training samples with K=50%	
		Features eliminated	Subset Features	I	H		I	H
AVIRIS Indiana Pine mini data set	Alfalfa	99	22	52%	12	6	66%	9
	Corn-notill	100	43	45%	13	5	70%	8
	Grass-trees	103	44	52%	11	3	69%	10
	Hay-windrowed	120	39	37%	12	5	71%	11
	Oats	98	59	21%	16	4	84%	19
	Soybean-mintill	97	38	56%	17	4	58%	10
	Woods	100	35	45%	12	3	67%	12
	Stone-Steel-Towers	121	52	46%	13	5	51%	15
ROSIS Pavia University mini data set	Asphalt	36	43	49%	07	3	67%	9
	Meadows	44	40	51%	12	2	71%	20
	Trees	39	51	55%	15	3	76%	10
	Painted metal sheets	56	48	58%	18	4	79%	12
	Bare Soil	48	42	48%	10	3	67%	9

Table 6. Classification results of different target class guided dimensionality reduction techniques.

Data Set	CoI	Total samples	FS		FS		FT	
			ρ	γ	ρ	γ	ρ	γ
AVIRIS Indiana Pine mini data set	Alfalfa	4	0	0	0	0	0.5	0.5
	Corn-notill	109	0.099	0.089	0.477	0.825	0.908	0.9
	Grass-trees	56	0.012	0.023	0.785	0.721	0.892	0.833
	Hay-windrowed	36	0.45	0.067	0.666	0.774	0.861	0.885
	Oats	3	0.00	0.00	0.00	0.00	0.333	0.2
	Soybean-mintill	188	0.101	0.156	0.627	0.776	0.957	0.923
	Woods	97	0.06	0.058	0.855	0.855	0.938	0.919
	Stone-Steel-Towers	7	0.01	0.125	0.142	0.111	0.714	0.5
ROSIS Pavia University mini data set	Asphalt	57	0.109	0.10	0.807	0.821	0.912	0.912
	Meadows	162	0.189	0.099	0.629	0.809	0.876	0.916
	Trees	26	0.05	0.033	0.538	0.451	0.846	0.758
	Painted metal sheets	12	0.009	0.004	0.50	0.345	0.56	0.62
	Bare Soil	43	0.03	0.078	0.57	0.456	0.61	0.64

IV. CONCLUSION

An experimental analysis is carried out to study the structural changes and class structures which resulted due to dimensionality reduction based on the control set of the target class. The philosophy of the target class guided feature selection and feature transformation is discussed followed by an experimentation to study the structural properties. A Minimum spanning tree is constructed at each stage of dimensionality reduction and the results are then compared. To study the neighborhood structure K-NN is also used and a comparative analysis has been presented. Further, for the purpose of studying the class structure agglomerative clustering is applied and the class structure is analyzed. Based on the structural analysis results, a new dimensionality reduction can be designed for preserving the structure that occurs upon reducing the feature as required by text, audio or video signals. The class structure analysis can help in optimizing the principal components which can be the basis of a future work.

REFERENCES

[1]. Shinn-Ying Ho , Chia-Cheng Liu, Soundy Liu, "Design of an optimal nearest neighbor classifier using an intelligent genetic algorithm", Pattern Recognition Letters(2002), vol 23, no 13, 1495–1503.

[2]. Daoqiang Zhang, Songcan Chen, Zhi-Hua Zhou, Learning the kernel parameters in kernel minimum distance classifier, Pattern Recognition(2006) , vol 39, no 1, 133–135.

[3]. Alexandros Iosifidis, Moncef Gabbouj., Multi-class Support Vector Machine classifiers using intrinsic and penalty graphs, Pattern Recognition (2016).

[4]. Dewan Md. Farid , Li Zhang , Chowdhury Mofizur Rahman , M.A. Hossain , Rebecca Strachan, "Hybrid decision tree and naive Bayes classifiers for multi-class classification tasks", Expert Systems with Applications(2014), vol 41 , 1937–1946.

[5]. Bartosz Krawczyk, Michał Woźniak , Gerald Schaefer,"Cost-sensitive decision tree ensembles for effective imbalanced classification", Applied Soft Computing(2014), vol 14, 554–562.

[6]. Yu-Lin He , Ran Wang , Sam Kwong, Xi-Zhao Wang, Bayesian classifiers based on probability density estimation and their applications to simultaneous fault diagnosis, Information Sciences(2014), Elsevier, vol 259, 252–268.

[7]. Alexandros Iosifidis , Moncef Gabbouj., Multi-class Support Vector Machine classifiers using intrinsic and penalty graphs, Pattern Recognition (2016) .

[8]. Sarah Vluymans , Danel Sanchez Tarrago, YvanSaeys, Chris Cornelis, Francisco Herrera., Fuzzy rough classifiers for class imbalanced multi-instance data, Pattern Recognition53 (2016), 36–45.

[9]. Patel, V.M. Gopalan, R. ; Ruonan Li ; Chellappa, R., "Visual Domain Adaptation: A survey of recent advances" , Signal Processing Magazine, IEEE.(2015) , vol 32 , no 3, 53 – 69.

[10]. Begum Demir, Sarp Erturk," Hyperspectral Image Classification Using Relevance Vector Machines", IEEE Geoscience and remote sensing letters(2007), vol. 4, no. 4, 586-590.

[11]. Marina Sokolova, Guy Lapalme, "A systematic analysis of performance measures for classification tasks", Information Processing & Management(2009), vol 45, no 4, 427–437.

[12]. Chao He , Mark Girolami, Gary Ross., Employing optimized combinations of one-class classifiers for automated currency validation, Pattern Recognition 37 (2004), 1085 – 1096

[13]. Jordi Munoz-Mari, Francesca Bovolo, Luis Gomez-Chova, Lorenzo Bruzzone Gustavo Camps-Valls, "Semi supervised One-Class Support Vector Machines for Classification of Remote Sensing Data", IEEE transactions on geoscience and remote sensing 8 (2010), vol. 48, 3188-3197

[14]. Dries F. Benoit, Rahim Alhamzawi,Keming Yu, "Bayesian lasso binary quantile regression", Computational Statistics(2013) , vol 28, no 6, 2861-2873.

[15]. Zhiquan Qi , Yingjie Tian' , Yong Shi, "Robust twin support vector machine for pattern classification", Pattern Recognition 1(2013), vol 46, 305–316.

[16]. Jurgen Schmidhuber, "Deep learning in neural networks: An overview", Neural Networks(2015), vol 61, 85–117.

[17]. Yuan-Hai Shao, Wei-Jie Chen, Nai-Yang Deng., "Nonparallel hyperplane support vector machine for binary classification problems", Information Sciences(2014), vol 263, 22–35.

[18]. Moya M., Koch, M., Hostler, L., One-class classifier networks for target recognition applications". Proceedings world Congress on Neural Networks, pp.797-801.

[19]. Young-Sik Choi , "Least squares one-class support vector machine", Pattern Recognition Letters (2009), vol 30, 1236–1240

[20]. The effective use of the one-class SVM classifier for hand written signature verification based on writer-independent parameters, Yasmine Guerbai, Youcef Chibani, Bilal Hadjadji., Pattern Recognition 48(2015), 103–113.

[21]. **David M.J.Tax, Robert P.W.Duin.,"Support Vector Data Description", Machine Learning1(2004), vol 54, no 1, 45-66.**

[22]. David M.J.Tax, Klaus-Robert M¨uller.," A Consistency-Based Model Selection for One-class Classification", International Conference on Pattern Recognition(2004), IEEE, vol 3, 363-366.

[23]. Shehroz S. Khan and Michael G. Madden, "A Survey of Recent Trends in One Class Classification", Springer-Verlang Berlin(2010), 188-197.

[24]. P.Nagabhushan, H.N.Meenakshi, "Target Class Supervised Feature Subsetting", International Journal of Computer Application(2014), vol 91, 12, 11-23.