Incremental Dimensionality Reduction in Hyperspectral Data

Preeti Mahadev
University of Mysore,
Mysuru, Karnataka,
India

P. Nagabhushan
University of Mysore,
Mysuru, Karnataka,
India

ABSTRACT

Conventionally, pattern recognition problems involve both samples and features that get collected over time or that gets generated from distributed sources. The system starts to falter when the number of features reaches a certain threshold and exhibits the curse of dimensionality. Traditionally dimensionality reduction (DR) is performed to prevent the curse of dimensionality when all features are available or when the system starts to degrade in its performance. But in the current digital age systems, the enormous and continuous influx of both samples and features mandates performing DR at regular intervals to keep up with the system performance. The massive amount of feature space and sample space that gets accumulated instantaneously allows little chance to extract the knowledge effectively that can be used promptly and hence mandates performing the DR at regular intervals of time. In real time scenarios, for any domain, decisions have to be made as and when the data is made available to realize the best outcome and to mitigate the risks. The various ways in which the features flow or get generated can be different depending on the domain of the dataspace. Due to its ever changing environment, extraction of knowledge can get more challenging. To overcome this problem of big data, an incremental dimensionality reduction (IDR) approach to extract, carry forward, build and accumulate the knowledge without recalling the previous data is explored in this case study. Both Feature subsetting and Feature transformation methods are employed for the purpose of illustrating the incremental reduction of attributes. The hyperspectral image generated from an AVIRIS sensor provides a versatile environment required to demonstrate the in depth study of an IDR approach. This case study attempts to showcase a novel approach of maximizing the knowledge while minimizing the information loss through the use of IDR techniques in a multifaceted environment with hyperspectral data.

General Terms

hyperspectral data, big feature space, sequence compulsive, sequence optional, correlation index

Keywords

Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), dimensionality reduction (DR), Incremental dimensionality reduction (IDR), Principal Component Analysis (PCA), Prims like approach, Kruskals like approach

1. INTRODUCTION

In a n-dimensional feature space where 'n' is a very large number, the sparsity of the data space will be very high [7]. As the volume of dataspace increases, the information to be gathered will be irregularly scattered. When sparsity increases, it mandates the reduction of features which will aid in the accumulation of the knowledge while increasing the compactness of the feature space. Instead of working with 'n' dimensions,'d' dimensions can be used where d<<n to reduce computational costs and to enhance visualization. There are two important methods of dimensionality reduction. The first one being Feature Subsetting (FS) and the second one being Feature transformation (FT) [2]. In Feature Subsetting, a subset of key features represent the entire feature space by retaining independent features and by discarding dependent features that are identified and selected using a filter approach or a wrapper approach [11, 12, 13]. The common aim of both approaches is to obtain the maximum knowledge by using minimum number of features. In the Filter method, a near optimal subset of features are selected based on a certain filter criterion or a rank. Filter method doesn't aim at evaluating the performance of the model and no learning is involved [2]. In a Wrapper approach, a near optimal subset is generated by learning and evaluating the model at each step of building the optimal subset. This subset is termed as an optimal subset because it represents the entire feature set by eliminating redundant features and by compacting the maximum knowledge. Knowledge can be in terms of variance of the feature set, classification accuracy, and predictor performance or decision making criterion. In Feature Transformation, the original feature space is transformed into another space of equal dimension. Usually, if the DR is performed in a supervised environment, Linear Discriminant Analysis (LDA) is employed and if the DR is performed in an unsupervised environment, Principal Component Analysis (PCA) is employed. To prevent the curse of dimensionality due to the huge influx of data, a customized and an incremental approach of the traditional DR method known as Incremental dimensionality reduction (IDR) is proposed. When a big dataspace is involved with a big, variant feature space and an invariant sample space, it becomes critical to explore the relationship among the attributes, extract knowledge as and when the information is gathered, without recalling the previously generated knowledge or previously processed data.

Remotely sensed multispectral data often involves large amount of data due to large spatial and multispectral coverage. Among the multispectral images, hyperspectral images is one such example that can generate a big feature space. Hyperspectral images are produced by instruments called imaging spectrometers. They contain a wealth of data, but interpreting them requires an understanding of exactly what properties of ground materials we are trying to measure, and how they relate to the measurements actually made by the hyperspectral sensor [5]. A typical example of hyperspectral image is the AVIRIS dataset acquired on June12, 1992 over the Purdue University Agronomy farm [1]. This scene was gathered by AVIRIS sensor over the Indian Pines test site in North-western Indiana and consists of 145\times145 pixels and 224 spectral reflectance bands in the wavelength range 0.4–2.5 10^(-6) meters[6]. This scene is a subset of a larger one. The Indian Pines scene contains two-thirds agriculture, and one-third forest or other natural perennial vegetation. There are two major dual lane highways, a rail line, as well as some low density housing, other built structures, and smaller roads. The scenario also pertains to some of the crops (corn and soybeans) being in their early stages of growth with less than 5% coverage. The ground truth available is designated into 16 classes of Vegetation and Buildings and are not all mutually exclusive. Indian pine data set captured from the AVIRIS sensor has each of its pixel represented by 220 spectral bands. The number of spectral bands have been reduced to 200 by removing bands covering the region of water absorption: [104-108], [150-163], 220 [1, 6].

Due to the small number of samples and the high number of features available in hyperspectral imaging applications, the need to reduce the attributes and to prevent the curse of dimensionality becomes inevitable [1-6, 10-13]. In order to maintain the performance of a classifier, performing DR at regular intervals in big feature space is mandated. But since the traditional DR process waits for all the features to get collected before reducing the feature space, an IDR approach is proposed to tackle the problem of the classier performance at regular intervals to enhance efficiency of the system. IDR aims at eliminating unproductive attributes, to accumulate the local knowledge and to progressively accumulate the maximum global knowledge.

2 INCREMENTAL FEATURE SUBSETTING WITH HYPERSPECTRAL DATA [FSIDR]

In this case study, an incremental filter approach based on proximity of attributes measured using correlation between attributes is used. A pairwise combination of all the possible attributes are considered and their correlation is calculated using PCC. For example, if a set of 200 attributes is considered, it can generate 39800 ($^{200}C_2$) correlations. Since the correlation of attributes is commutative in nature, one has to ideally calculate 19900($^{200}C_2$ / 2) correlations. Instead of working on an exhaustive list that gets enumerated while working with traditional FS, an incremental feature sub setting (FSIDR) approach is thought of to reduce the computational cost and to build the knowledge incrementally as and when the features are made available. At a given instance, only a selected set of features can be considered for the purpose of calculation, thus achieving parallelism. The nature of incremental flow of features can vary depending on the arrival/generation of features in the given environment [17, 18]. In this study, two scenarios are considered: One being analogous to the temporal flow of features which involves sequence compulsive arrival of features and the other one being analogous to features coming from distributed environment or from a multisensory environment which involves the generation of features that is sequence optional [15].

The framework for the incremental feature subsetting consists of two significant variables: Pearson's correlation coefficient (PCC) represented by 'r' and the threshold factor represented by 't'. The strength of the correlation measured between two features is quantified by 'r'. Threshold factor 't' is the cutoff value that is used to determine whether both the features considered are eligible to be retained or if one of the feature can be eliminated from the feature subset. The value of 't' is suggested to be 0.6 and has been utilized in the experiments of this study [17]. The PCC for every pair of attributes in the first batch is calculated. If the PCC value for a pair of attributes X, Y is less than 0.6, then both the attributes are added to the optimal subset because the attributes are considered to be uncorrelated to each other. If the PCC value is greater than 0.6, it indicates that both the attributes are highly correlated and the presence of both in the optimal subset is deemed redundant. So the deciding criteria to retain one of them will be the attribute that has a higher standard deviation. Higher standard deviation in an attribute indicates that the variance is higher and hence gets added to the optimal subset that is being built while deleting the one with a lower standard deviation.

2.1 Sequence Compulsive model for FSIDR

Let us consider the AVIRIS dataset, a hyperspectral image of Indian Pines region in Indiana in 1992[1-6]. Using target class guided compression in Feature subspace, Meenakshi et al have extracted the spectral signature of each class individually and the corresponding sample space for each of the 16 classes [16]. In hyperspectral data, due to the high possibility of pixels getting mixed up in the presence of the overlapping spectral bands, two different datasets have been derived out of the AVIRIS data for the purpose of exploration. Using the spectral signatures and the corresponding sample set for the classes of interest and to illustrate the IDR in high dimensional spaces, the two subsets, also referred to as the AVIRIS mini datasets namely overlapping class dataset and distinct class dataset are used. A set of 5 overlapping classes and another set of 3 distinct classes are considered for realizing the trends and patterns while performing IDR. The overlapping classes dataset consists of 1737 samples comprising of 5 classes namely: alfalfa, corn mintill, soy clean, stone tower and wheat. Since the corn and soya samplings are at their early stages of development, they have not yet developed the discriminating features that are distinctive in nature. Hence the spectral signature for the 2 classes overlap extensively during classification. The distinct class data consist of 344 samples with 3 classes namely: alfalfa, wheat and stonetower. This dataset has distinctive features that differentiates the clusters better as their spectral signatures don't overlap as much as the overlapping class dataset.

Table 1: Optimal subset of distinct class dataset for temporal FSIDR

Batch	Attributes for DR	Optimal Subset
{B1}	{B1}	{A, B}
{B1} + {B2}	{A, B} + {B2}	{A, C, D}
{B1} + {B2} + {B3}	{A, C, D} + {B3}	{A, C, D, E, F}
{B1} + {B2} + {B3} + {B4}	{A, C, D, E, F} + {B4}	{A, C, D, E, F}
{B1} + {B2} + {B3} + {B4} + { B5}	{A, C, D, E, F} + {B5}	{A, C, D, E, F, X}

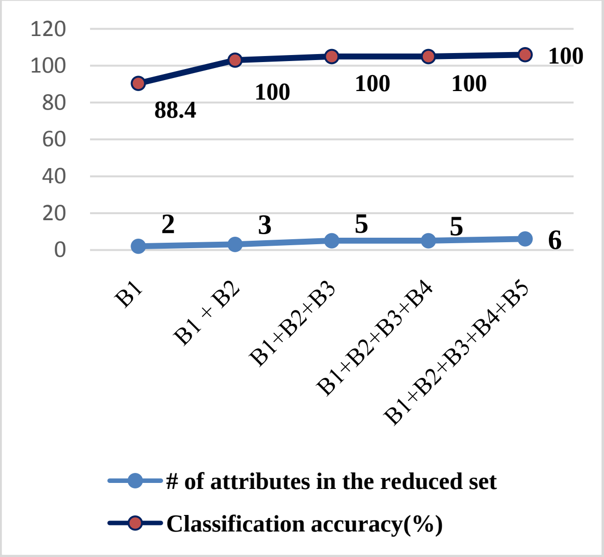


Fig 1: Temporal FSIDR of distinct class dataset

Let us assume that the features from the distinct dataset are arriving temporally over time, 40 features in each batch with 5 batches i.e. batches B1, B2, B3, B4 and B5. On applying incremental feature subsetting procedure, the first batch of 40 features reduces itself to an optimal subset of 2 features { A, B} and provides 88.4% classification accuracy [see Figure 1]. These two features from B1 is added to the next 40 features of B2 and is incrementally reduced to {A, C, D}. Similar procedure is carried out for batches B3, B4 and B5 [see Table 1]. At the end of FSIDR, the optimal feature subset consists of 6 features {A, C, D, E, F, X} and achieves 100% classification accuracy. The same results are obtained when traditional FS is performed with all features reduced at once. This supports the hypothesis made earlier that sub optimal knowledge can be obtained incrementally in the interim and an optimal knowledge can be progressively built without waiting for all the features to arrive. This in turn bolsters the decision making process because the interim knowledge generated incrementally is not only synchronous to the knowledge that is made available at the end but is also in line to the knowledge extracted when DR is performed with all the features at once. At any instant of time, the decision making criterion based on incremental learning as illustrated in this section will be built upon the unseen data of the future, without recalling the previous data.

2.2 Sequence Optional model for FSIDR

Features can be generated from different sources or from multisensors as in features generating from a video surveillance with video cameras focusing from different directions. Traditionally, the features need to merged centrally in order to extract the global knowledge. So, data is collected in one central repository and once all features are collected, the knowledge extraction is performed. This would require a prolonged waiting time, computationally intensive operation due to the massive amount of data being dealt at once etc. To avoid these problems, the IDR process proposes to reduce the data as and when it is generated and build the overall knowledge as it incrementally merges the previously generated knowledge with the new set of features that is being generated.

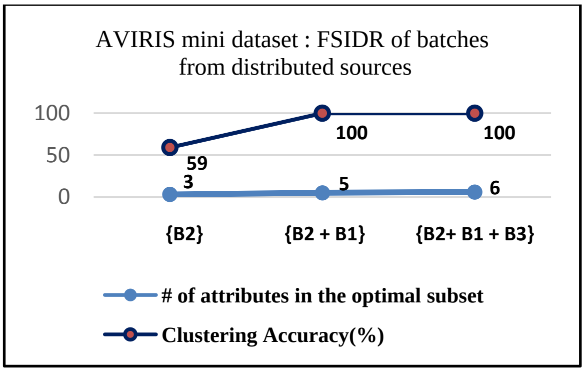


Fig 2: Distributed FSIDR of distinct class dataset

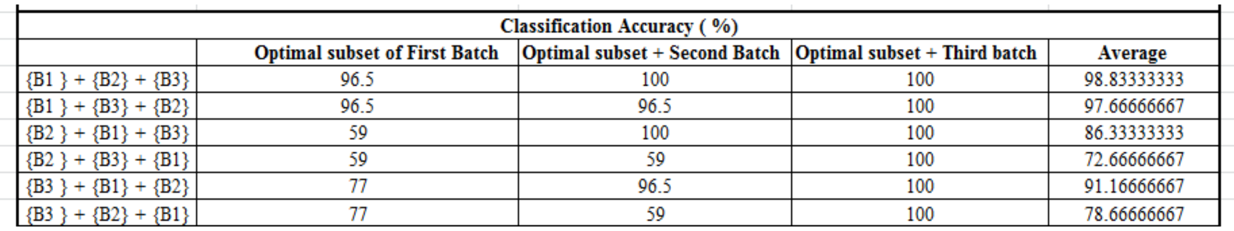
Let us assume that the AVIRIS minidataset distinct class data is generated from 3 different sources: Batch 1, Batch2 and Batch 3 to simulate a multisensory environment. The 200 features of the dataset is divided into 3 batches, chosen heuristically for the purpose of exploration. Batch 1 consists of features 1 to 74 ; Batch 2 consists of features 75 to 159 and Batch 3 consists of features 160 to 200.For a given set of 3 batches there can be 6 (3!) different sequences in which they can be merged (see Table 2). For illustration let us consider the sequence {B2, B1, B3}. Batch 2 represented by B2, generates an optimal subset of 3 features with a classification accuracy of 59%. These 3 features from Batch 2 are incrementally merged with Batch1 to obtain 5 features in the optimal subset resulting in 100% classification accuracy (see Figure 2).

The third batch is also incrementally merged in a similar way and the classification accuracy remains at its maximum even after adding an additional feature to the optimal subset from Batch 3. The experiments carried out illustrate that the knowledge will either be carried forward progressively or will remain at its maximum thus bolstering the FSIDR approach of merging the batches in a distributed environment. The order of merging of batches are interchanged and the experiments are carried out as explained. Although the same results were obtained at the end of merging all the batches, the interim knowledge obtained may vary from sequence to sequence (see Table 2). In this example with 3 batches, Batch1 has 96.5% , Batch2 has 77% and Batch 3 has 59% individual classification accuracy. Batch 1, which has the least misclassification, when merged earlier in the merging sequence gives a better average classification accuracy. If Batch 3 which has the highest misclassification, when merged earlier provides the least average classification accuracy. Generally it is be observed that in AVIRIS data, if batches with less misclassification are merged earlier, the subsequent merges tend to converge towards the right classification as it moves further incrementally.

The results of the experiments with 3 batches shows that all the merging sequences achieve 100% classification at the end of the IDR (see Table 2). It can also be observed that the interim classification accuracy varies from sequence to sequence in the order they are merged and reduced. The merging sequence {B1, B2, B3} is the most optimal one because it achieves the highest average classification accuracy, i.e. 98.8%. (see Figure 3). It implies that choosing the optimal merging sequence not only provides maximum classification accuracy at the end but also provides the best interim knowledge throughout the IDR process.

It is observed that in this optimal merging sequence, the spectral values are arranged in the descending order .The average spectral value of B1, B2 and B3 are 4314, 1754 and 1126 respectively. Conversely it is also observed that if the batches with lower spectral values are merged earlier as in {B3, B2, B1} and {B2, B3, B1}, the average classification accuracy dips to a lowest level of average classification accuracy.

Table 2 : Phasewise classification Accuracy of all the combinations of the 3 batches using distinct dataset



	Classification Accuracy (%)			
	Optimal subset of First Batch	Optimal subset + Second Batch	Optimal subset + Third batch	Average
{B1 } + {B2} + {B3}	96.5	100	100	98.83333333
{B1 } + {B3} + {B2}	96.5	96.5	100	97.66666667
{B2 } + {B1} + {B3}	59	100	100	86.33333333
{B2 } + {B3} + {B1}	59	59	100	72.66666667
{B3 } + {B1} + {B2}	77	96.5	100	91.16666667
{B3 } + {B2} + {B1}	77	59	100	78.66666667

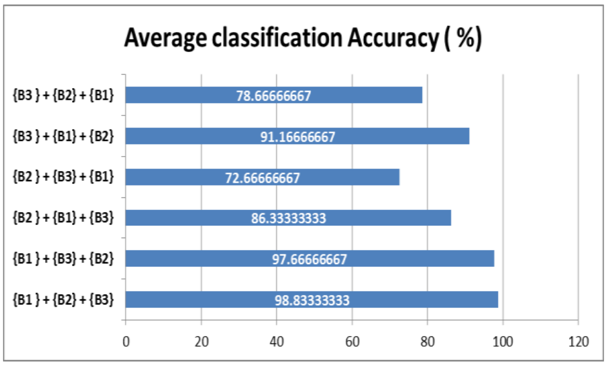


Fig 3 : Average Classification Accuracy of the various combinations of batches of distinct dataset

To corroborate the deductions of the above experiment, the same dataset is divided into 8 batches B1…B8 with 25 features each in sequence for further experimentation. The average spectral values of each batch is calculated and is listed in the descending order of the spectral values. With 8 batches, there can be 40320 (8!) merging sequences that can be generated (see Table 3). But since it is seen in previous section by using an exhaustive set that the decreasing order of spectral values generates the best average classification accuracy, two sequences i.e. descending order of batches and ascending order of batches in terms of their average spectral values are considered (see Table 3).

Table 3: Average Spectral values of the 8 batches of the distinct dataset

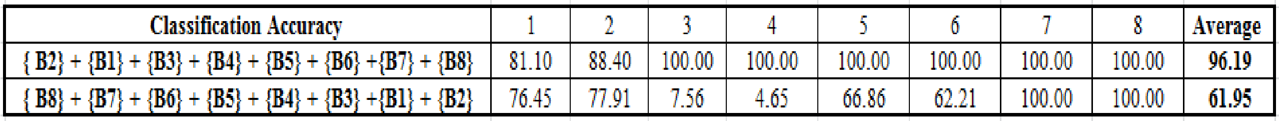
Batch	Average Spectral value
B2	4569.61
B1	4366.66
B3	3952.62
B4	2496.48
B5	1513.92
B6	1425.93
B7	1179.37
B8	1080.51

The results of merging 8 batches incrementally is as shown (see Table 4). It can be observed that if the batches with higher spectral values are merged earlier i.e. the merging sequence { B2} + {B1} + {B3} + {B4} + {B5} + {B6} +{B7} + {B8} generates a consistent and progressive knowledge during its course of IDR and knowledge extraction. The converse holds true as well as the second sequence { B8} + {B7} + {B6} + {B5} + {B4} + {B3} +{B1} + {B2}, arranged in the descending order of average spectral bands generates approximately 30% less classification accuracy during the course of IDR. Since IDR can provide the best decision making criteria at any given point of time, it is critical to identify the optimal merging sequence and it is suggested that the merging of batches should follow the rule of merging higher spectral values earlier to achieve best results .

3. INCREMENTAL FEATURE TRANSFORMATION WITH HYPERSPECTRAL DATA [FTIDR]

In feature subsetting, although redundant features are eliminated to obtain the reduced space, eliminating the features in its entirety poses a risk of information loss. An alternative method of DR known as feature transformation is considered that does not eliminate any features but recreates a transformed feature space from the original feature space. The transformed feature space comprises of a fraction of each of the original features by assigning a certain weightage to each of the feature. [22].

Table 4 : Phase wise (FTIDR) classification accuracy of the 8 batches of the distinct dataset



Classification Accuracy	1	2	3	4	5	6	7	8	Average
{ B2} + {B1} + {B3} + {B4} + {B5} + {B6} +{B7} + {B8}	81.10	88.40	100.00	100.00	100.00	100.00	100.00	100.00	**96.19**
{ B8} + {B7} + {B6} + {B5} + {B4} + {B3} +{B1} + {B2}	76.45	77.91	7.56	4.65	66.86	62.21	100.00	100.00	**61.95**

3.1 Sequence compulsive model for FTIDR using hyperspectral data

Hyperspectral data characteristically comprises of several features that define a given pixel in an unsupervised environment [1-6]. To reduce the feature space, IDR is performed by employing Principal Component Analysis (PCA). The variability of the principal components (PCs) aid to determine the number of PCs to be considered significant enough for a given model.

In our previous studies where datasets like Dow Jones data, Corn Soya data, and Student datasets have been used, 90% cumulative variance is considered as the threshold to classify the data effectively [17,18]. In the AVIRIS dataset since the spectral values have high variability throughout the dataset and the classes are not mutually exclusive like the usual datasets, the PCs hold significant variance even beyond 90% cumulative variance. Hence by trial and error, the point where the scree plot starts to straighten is determined to be a quantitatively reasonable threshold to classify the dataset effectively in hyperspectral data. For example if the percentage of cumulative variances for a given dataset are 82.6, 94.1, 97.5, 98.3, 98.9, 99.4, 99.5, 99.6, 99.7 … 99.9 and 100 respectively, the variability of the first 6 components will be considered as most significant because the % variability starts to stabilize at the threshold (see Figure 4). In this study, the knowledge obtained is quantitatively measured in terms of classification accuracy. In this example the classification accuracy achieved by using the 6 PCs with cumulative variance of 99.4 % doesn't improve by adding additional PCs. It indicates that after the threshold point, the variability of the PC is minimal or negligible and this in turn implies that it would not add significantly to the knowledge that is being incrementally accumulated.

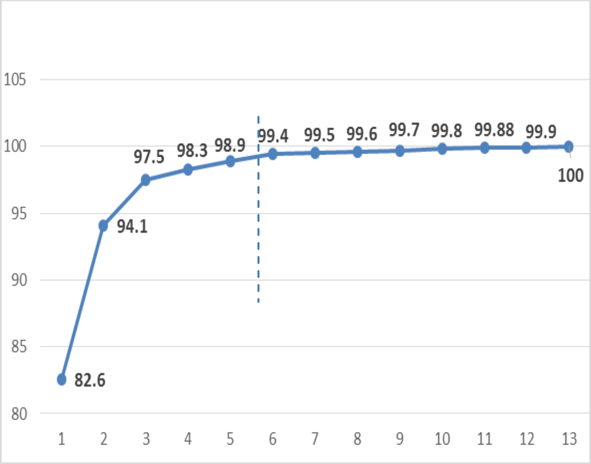


Fig 4 : Scree plot for choosing the number of principal components

3.1.1 Sequence Compulsive model for FTIDR using distinct dataset

While performing the incremental merging of the features assumed to be arriving temporally, two different approaches were considered [18]. The first approach will be to incrementally merge the PCs successively. The second approach will be to incrementally merge the original features successively instead. Both the approaches are applied to the distinct class dataset and results are as shown (see Figure 5, 6, 7). The trend in which the pattern changes as the batches are merged in both approaches remains visually identical. Quantitatively, although the classification accuracy obtained and the cumulative variance accumulated in the second approach is slightly higher, the number of PCs used to achieve the classification is lesser in the first approach because the DR is performed on the PCs (see Figure 7). But for the first approach there is an additional step of converting the original attributes to PCs before merging can take place. Approaches are highly comparable: the second approach proves to be more efficient in terms of time and complexity of the algorithm.

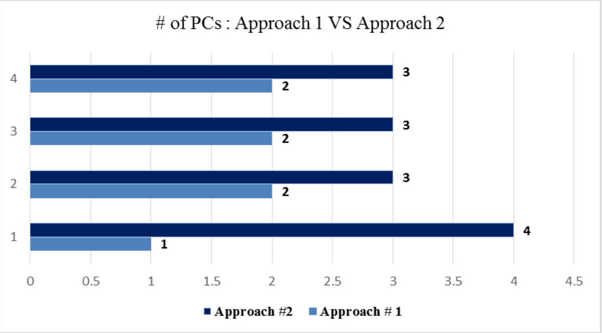


Fig 5: Comparison of classification accuracy required for the two approaches using distinct dataset

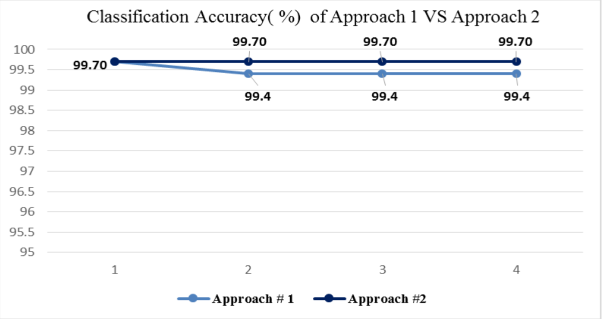


Fig 6 : Comparison of PCs required for the two approaches using distinct dataset

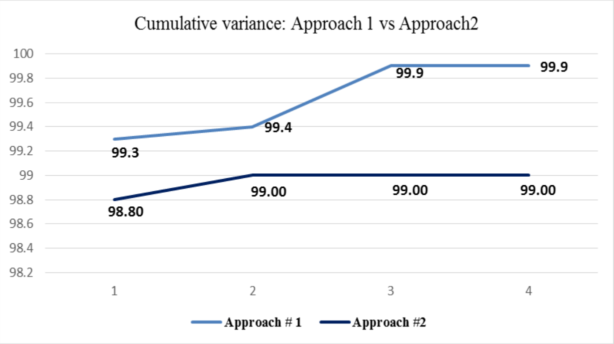


Fig 7: Comparison of cumulative variance accumulated phasewise for the two approaches using distinct dataset

3.1.2 Sequence Compulsive model for FTIDR using overlapping dataset

The overlapping class dataset with 5 classes and 200 features for each sample is assumed to arrive with 25 features in each batch (see Figure 8). Hence 8 batches with 25 features in each batch are assumed to arrive sequentially. As the features arrive temporally, the change in trend is visually irregular and can be observed from one batch to another which is due to the presence of vastly varying spectral ranges in each batch .As shown in the previous section, merging the original attributes instead of the PCs incrementally for performing DR is a computationally effective approach. So the same approach is used for illustrating the incremental merge of overlapping classes dataset as well. While incrementally merging the batches in this dataset, Corn and Soya classes are misclassified most of the time as they have overlapping spectral bands owing to its early developmental stages (see Figure 9). The other 3 classes have distinct spectral bands and are classified distinctively.

It can be observed that, although individual batches display irregular trend locally, the incremental merge consistently maintains the optimal pattern and generates progressive knowledge while converging towards the maximum possible knowledge globally (see Figure 8, 9). The incremental merge achieves approximately 60% classification accuracy at the end which corroborates the previous studies [11, 12]. The 60% classification accuracy that is achieved incrementally is the same when all the features are reduced together at once (see Figure 10). IDR not only projects an advantage over time but also supports the concept of availing the best possible cumulative knowledge at any instance of time for making well informed decisions.

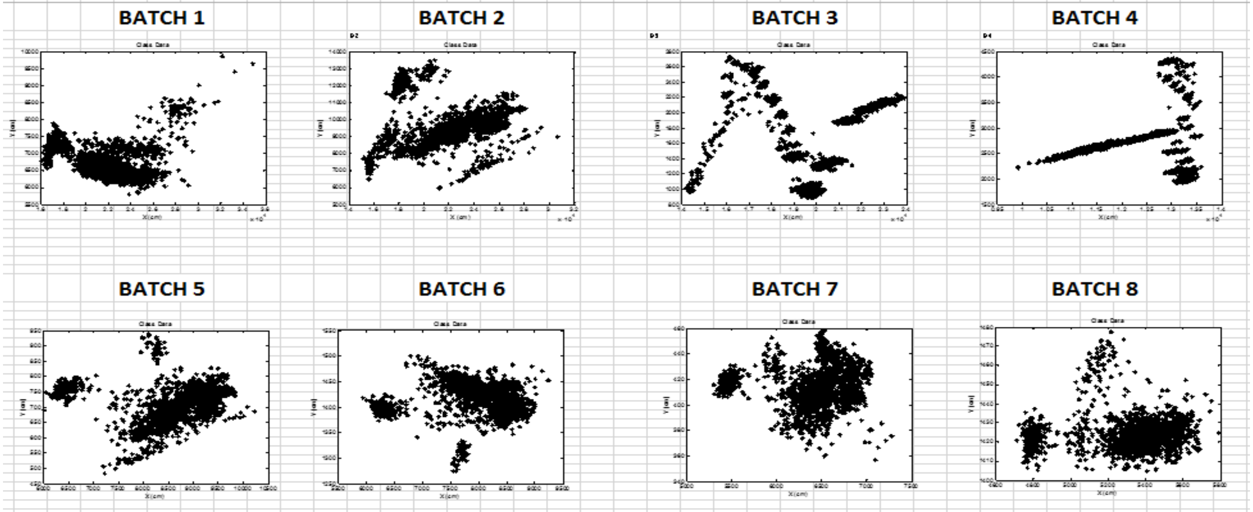


Fig 8: Local pattern in each batch of the overlapping dataset

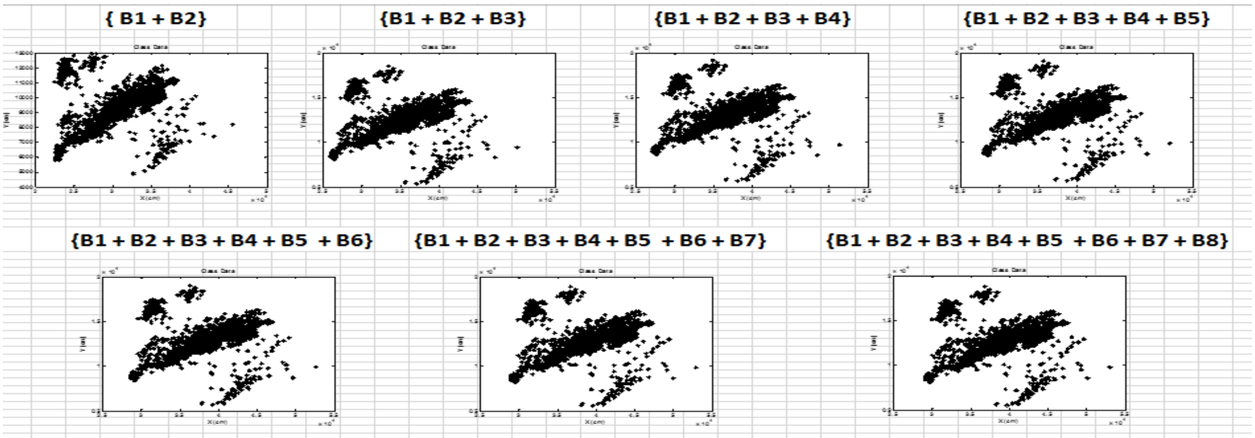


Fig 9: Global pattern accumulating at each incremental phase of the overlapping dataset

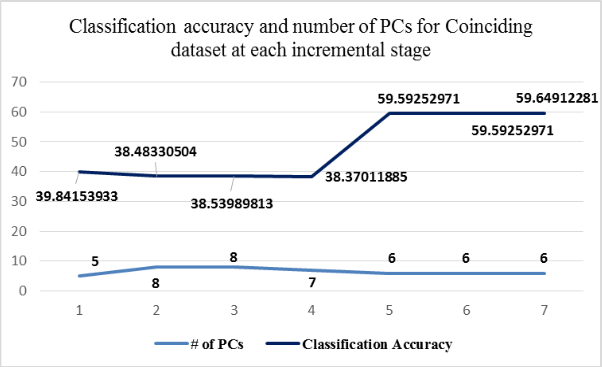


Fig 10: Classification accuracy and number of PCs for overlapping dataset at each incremental stage

3.2 Sequence optional model for FTIDR using hyperspectral data

In hyperspectral data where the number of samples is comparatively less than the number of features, the need for DR is more imminent [9-12]. Although hyperspectral data can be inherently thought of as sequence compulsive flow of features, one can simulate the data as is coming from the multisensors. Let us consider the distinct Class dataset of 3 classes to be generating from 5 different sensors with 40 features each (see Figure 12). The challenge will be to merge the features such that the best possible knowledge is accumulated locally and maximum knowledge is built globally with minimum loss of information. In an incremental mode, IDR will aim at gathering the interim knowledge progressively, accumulating maximum knowledge gradually, without recalling the previously processed data. Since the sequence of merging the features in a distributed environment in not sequence compulsive, identifying the order of merging the features that can achieve the maximum classification accuracy needs an organized approach. To arrive at the optimal merging sequence in a distributed environment, a framework based on the correlation based proximity measure and principal component is devised (see Figure 11).

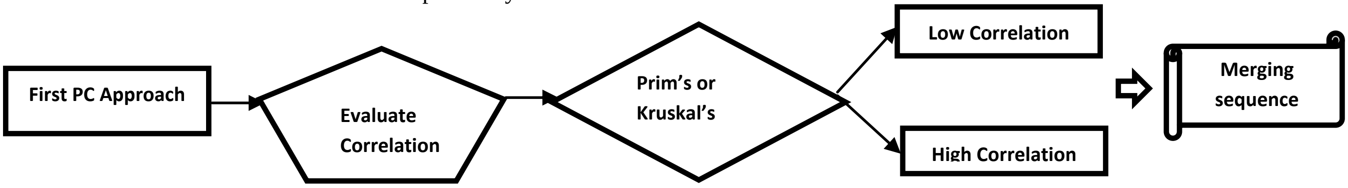


Fig 11 : Framework to identify the optimal merging sequence

The idea of PCA is achieved by transforming the original set of variables to a new set of variables known as principal components (PCs). The PCs will be uncorrelated to each other and the maximum variance of the feature set will be accumulated in the first few PCs [14,22]. Usually, the first PC itself will accumulate around 80% of the variance of the entire feature set. It is well established that the first PC will accumulate the maximum variance of the feature set [22]. Hence in our study, the first PC of each batch is considered to represent the corresponding batch. The correlation of all combinations of batches are calculated using only the first PC and is referred to as the first PC matrix. Since the distinct dataset has mutually exclusive classes, the same has been considered for illustration. For example, consider 5 batches of 40 features each from the distinct class dataset: B1, B2, B3, B4 and B5, the correlation of the possible combination of batches are calculated to generate the first PC matrix (see Table 5)

Table 5: Correlation index matrix of the distinct dataset with 5 batches

	B1	B2	B3	B4	B5
B1		-0.397	0.546	0.778	0.890
B2	-0.397		0.485	0.177	-0.058
B3	0.546	0.485		0.939	0.810
B4	0.778	0.177	0.939		0.959
B5	0.890	-0.058	0.810	0.959	

If the average PCC of the matrix is low i.e. <= 0.6, it indicates that the batches have high variance amongst themselves. In this case the average PCC =0.51 indicates that the uncorrelated batches must be merged successively to accumulate maximum knowledge progressively. On the other hand, if the average PCC is high (i.e. >=0.6), then the global knowledge can be built progressively by merging the highly correlated batches first. For a given 'n' batches, there are n! ways of generating the merge sequences.

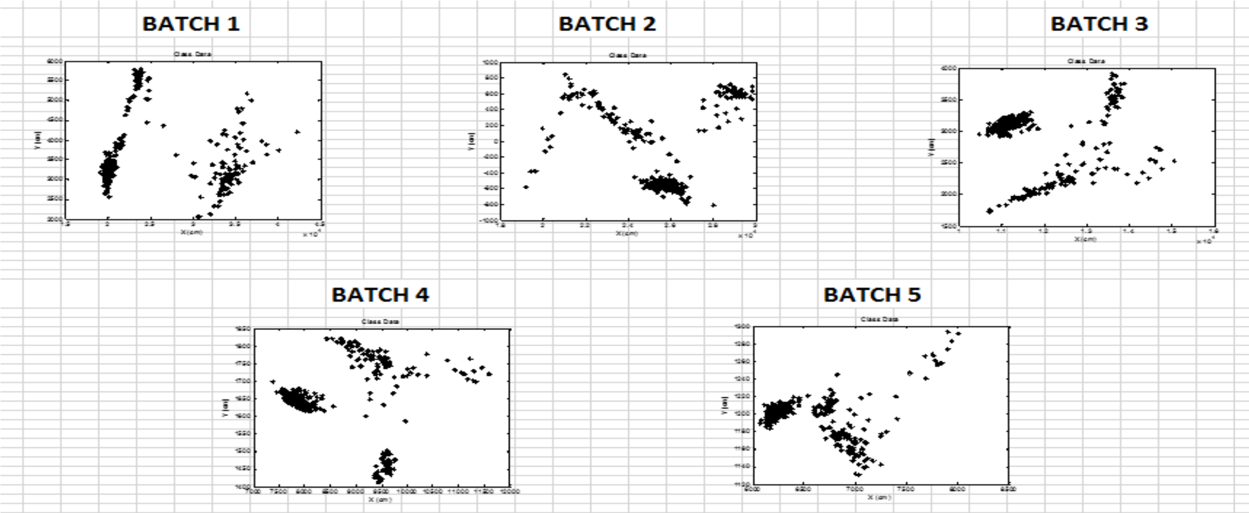


Fig 12: Local pattern in each batch of the distinct dataset

To identify the most optimal merge sequence that not only build knowledge progressively but also compacts maximum knowledge, either a Prims like approach or a Kruskal's like approach is followed. The correlation of the first PC matrix is arranged in a sorted order (see Table 6). For the purpose of illustration let us consider the Prims like approach with high correlation as criterion for ranking. The starting batch can be an arbitrary one in Prims like approach. For the ease of explaining, the starting batches for merging is considered to be B4 and B5. Now the optimal sequence will be {B4, B5}.The next in the list is {B3, B4}. Since B4 is already in the sequence, only B3 is selected for the merge arriving at {B4, B5, B3}. Similarly when the B1,B5 are considered the sequence now grows to {B4,B5,B3,B1} and finally end up as {B4,B5,B3,B1,B2} (see Figure 13).

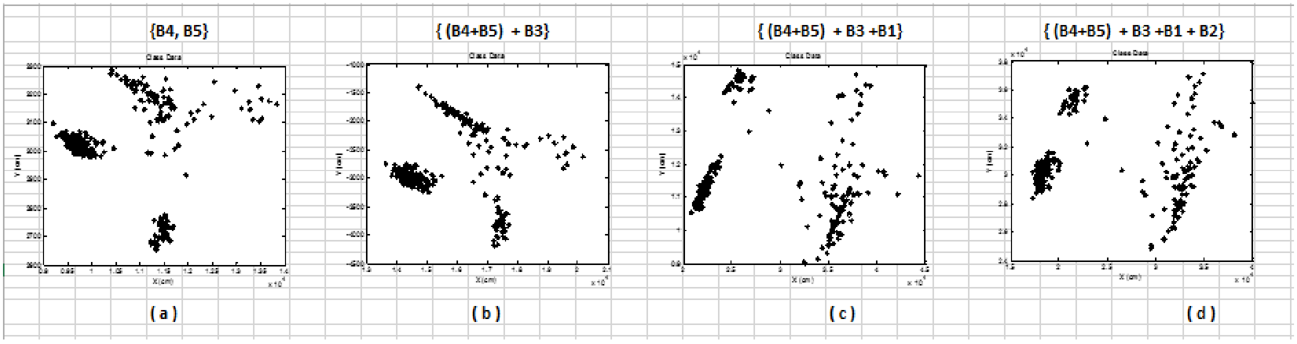


Fig 13: Phase wise merging pattern obtained for the optimal merging sequence with high correlation as priorit

On similar lines if low priority is considered as the criterion for ranking then we get the optimal sequence list as {B1, B2, B5, B4, B3} (see Figure 14). In Kruskals like approach, the first batch to start the merging sequence is fixed. The batch that satisfies the ranking criterion should be selected first. At every step, the batch need not be connected to the already selected batches in the optimal merging sequence as in the case of Prims like approach. Coincidentally for this dataset both Prim's like and Kruskal's like approach with low correlation as criterion for ranking will generate the same sequence of merging i.e. {B1, B2, B5, B4, B3}. Since the overall correlation value <0.6, it can be observed that the pattern of merging the batches with low correlation is more consistent with the final results than the pattern of merging the batches with high correlation successively (see Figure 13 and Figure 14).

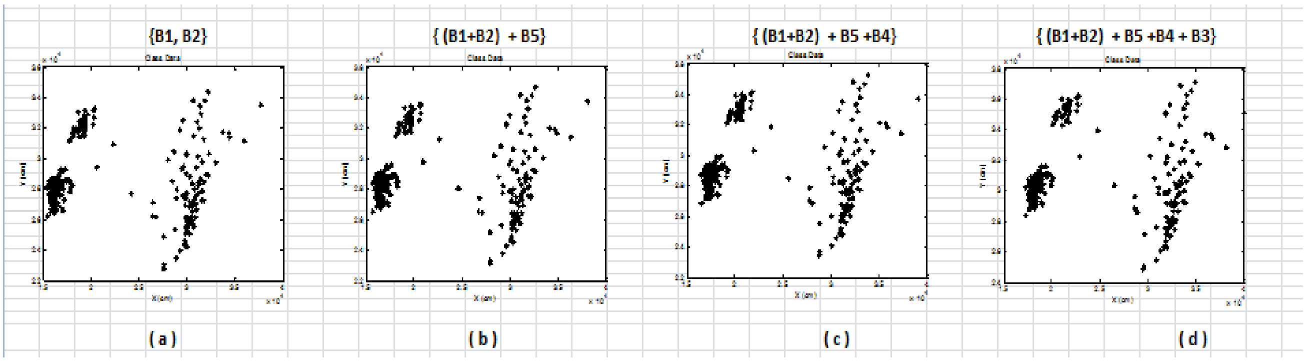


Fig 14: Phase wise merging pattern obtained for the optimal merging sequence with low correlation as priority

Table 6 : Sorted PCC values of the combinations of features using distinct dataset with 5 batches

B4B5	0.959
B3B4	0.939
B1B5	0.890
B3B5	0.810
B1B4	0.778
B1B3	0.546
B2B3	0.485
B2B4	0.177
B2B5	-0.058
B1B2	-0.397

For this dataset, if Prims like approach with high correlation as criterion is used for merging, the average classification remains low but if low correlation is considered then the average classification accuracy remains high. (see Figure 15). Prims like approach with low correlation also requires less number of PCS to achieve the accuracy and remains fairly consistent throughout when compared to Prims like approach with high correlation (see Figure 16). The cumulative variance accumulated at each phase for both high correlation and low correlation seems highly comparable (see Figure 17). This shows that when the overall variance of the dataset is low, merging the batches with lower correlation will generate more knowledge incrementally.

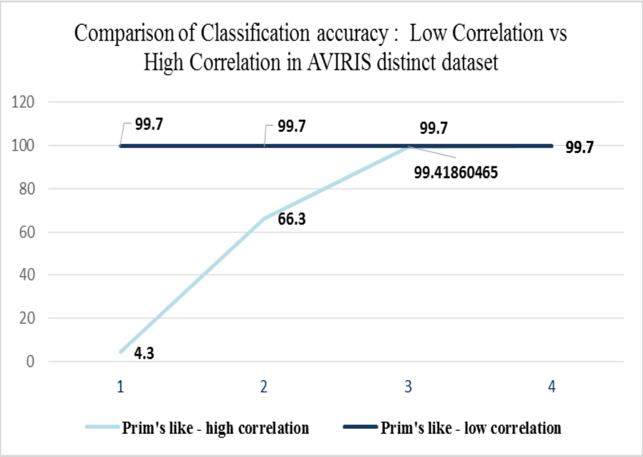


Fig 15 : Comparison of Classification accuracy : Low Correlation vs High Correlation in distinct dataset

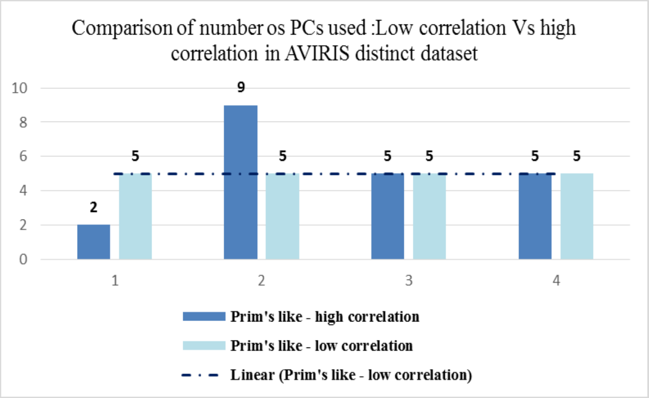


Fig 16: Comparison of number of PCs used :Low correlation Vs high correlation in AVIRIS distinct dataset

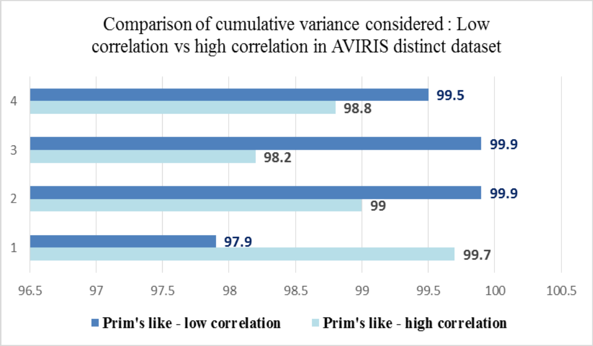


Fig 17 : Comparison of cumulative variance : Low correlation vs high correlation in AVIRIS distinct dataset

4. DIVIDE AND CONQUER PARADIGM AS A PROOF OF CONCEPT

The IDR method draws a certain analogy with the Divide and Conquer algorithmic paradigm on the lines of parallelism. Divide and Conquer (DnC) is an algorithmic paradigm that bears resemblance to dynamic programming and Greedy algorithms. In DnC, a given problem is broken down into sub problems[24]. Each sub problem is broken down into further sub problems until the final solution is reached. Usually in an ideal dataset, the number of samples will be significantly higher than the number of attributes. But in high dimensional datasets with big feature space, the number of attributes will be unusually higher than the number of samples as in hyperspectral data. Hence the feature space can be broken down iteratively to batches with heuristically sizeable features as done in DnC. In Summary, DnC procedure divides the problem into subproblems, reduces and merges the subproblems recursively and finally conquers the solution of the problem .The Divide and Conquer technique can be thought of as a proof of concept that demonstrates the parallelism achieved while performing IDR. It is also used to illustrate the advantages of IDR over DnC. The distinct dataset, a minidataset of AVIRIS is subjected to DnC using FS and FT techniques separately as follows.

4.1 Divide and Conquer using Feature Subsetting

For DnC using FS or FT, the 200 features are first broken down recursively until the features are considered to be of a manageable size i.e. approximately 12 features in a batch. (see Figure 18). The features are recursively broken down from Step #1 to Step #5 likewise. At Step #6, feature reduction starts to takes place. Hence, the steps for both the techniques are common until Step #5, which involves dividing the batches of features recursively (see Fig 18).

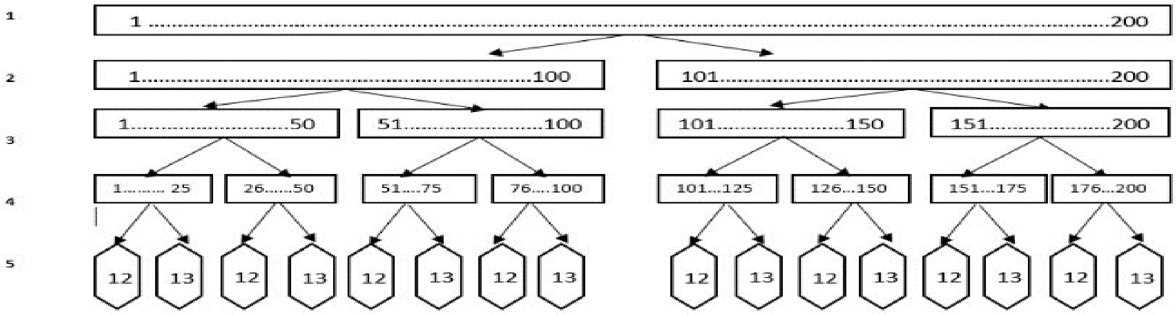


Figure 18 : Common steps for Divide and Conquer using FS and FT

For DnC using the FS technique, with a PCC threshold of 0.6 (60%) , two attributes are considered to have a high correlation in the given feature space [17]. The batches of features are reduced individually to get the reduced subset. At step#9, the reduced features are merged in pairs (see Figure 19).At Step #10, the merged features are reduced further. It can be observed that less number of features are obtained every time after reduction indicating that the redundant features are eliminated. At Step #11, the first 100 features are reduced to 5 features and the last 100 features are reduced to 4 features using the DNC procedure. These features are finally reduced to 6 features. At this point, the features cannot be reduced any further and hence reaches the final reduced feature space. The 6 features obtained this way provides 100% classification accuracy.

4.2 Divide and Conquer using Feature transformation

DnC with FT technique is similar to FS except that FT transforms the original feature space before reduction. PCA is employed here for illustration of the DR procedure. At step#6 after the batches of features are broken down to attain a manageable size of features, PCA is applied to each batch and the original feature space of 12 and 13 features are transformed into an eigen transformed feature space of the same number of features (see Figure 20). For reduction of the transformed feature space, the threshold point where the scree plot starts to slow down the variance is considered and is reduced accordingly. At step #8 , the reduced features are merged and since the merging can introduce redundancy, the merged features are reduced again. This process of merge, reduce and conquer is carried out recursively and at step #14, only six features are obtained .These 6 features are further reduced to get the final solution of 4 features that represent the entire feature space. At this point the feature space cannot be reduced any further.

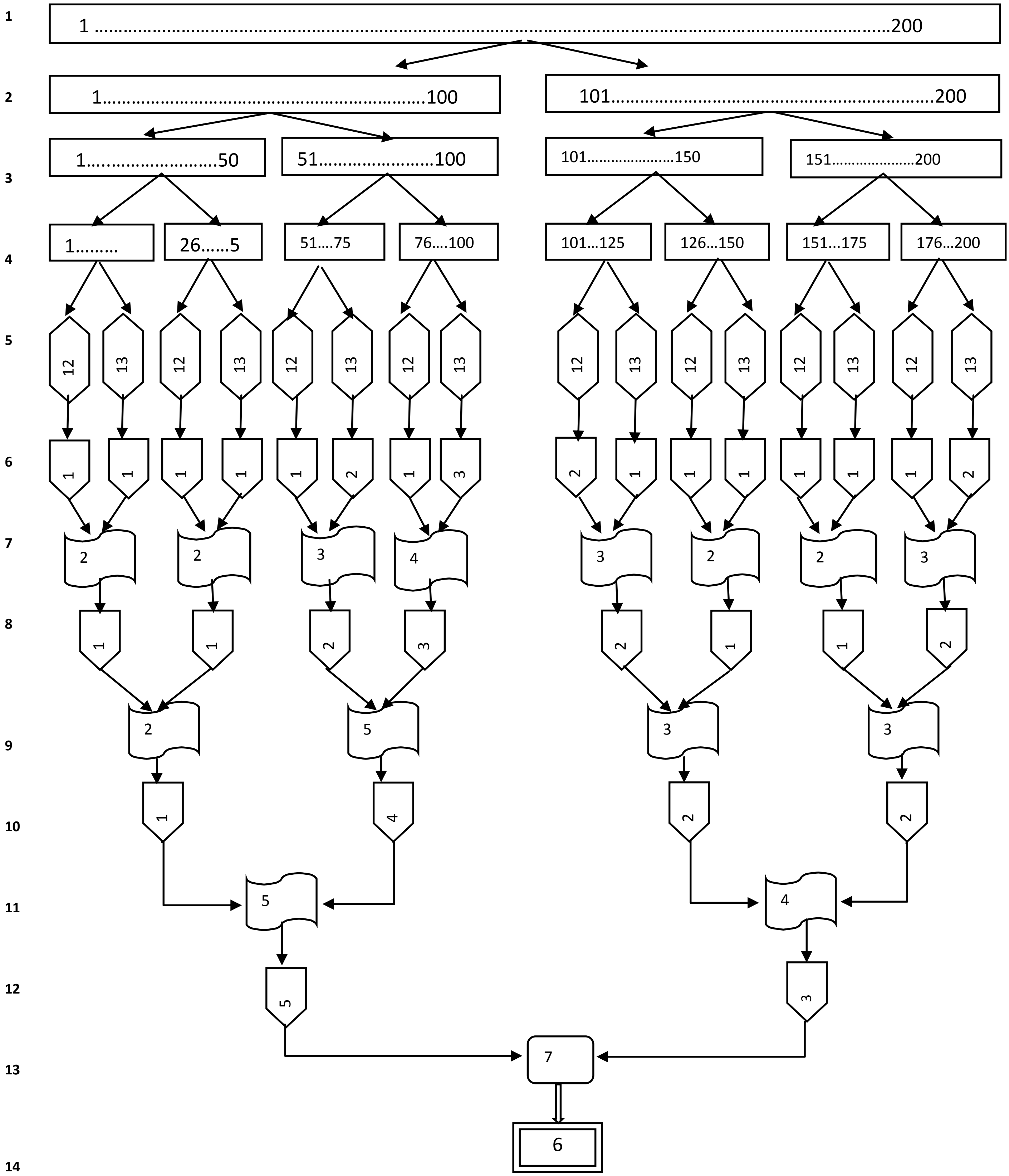


Figure 19: Divide and Conquer using Feature Subsetting

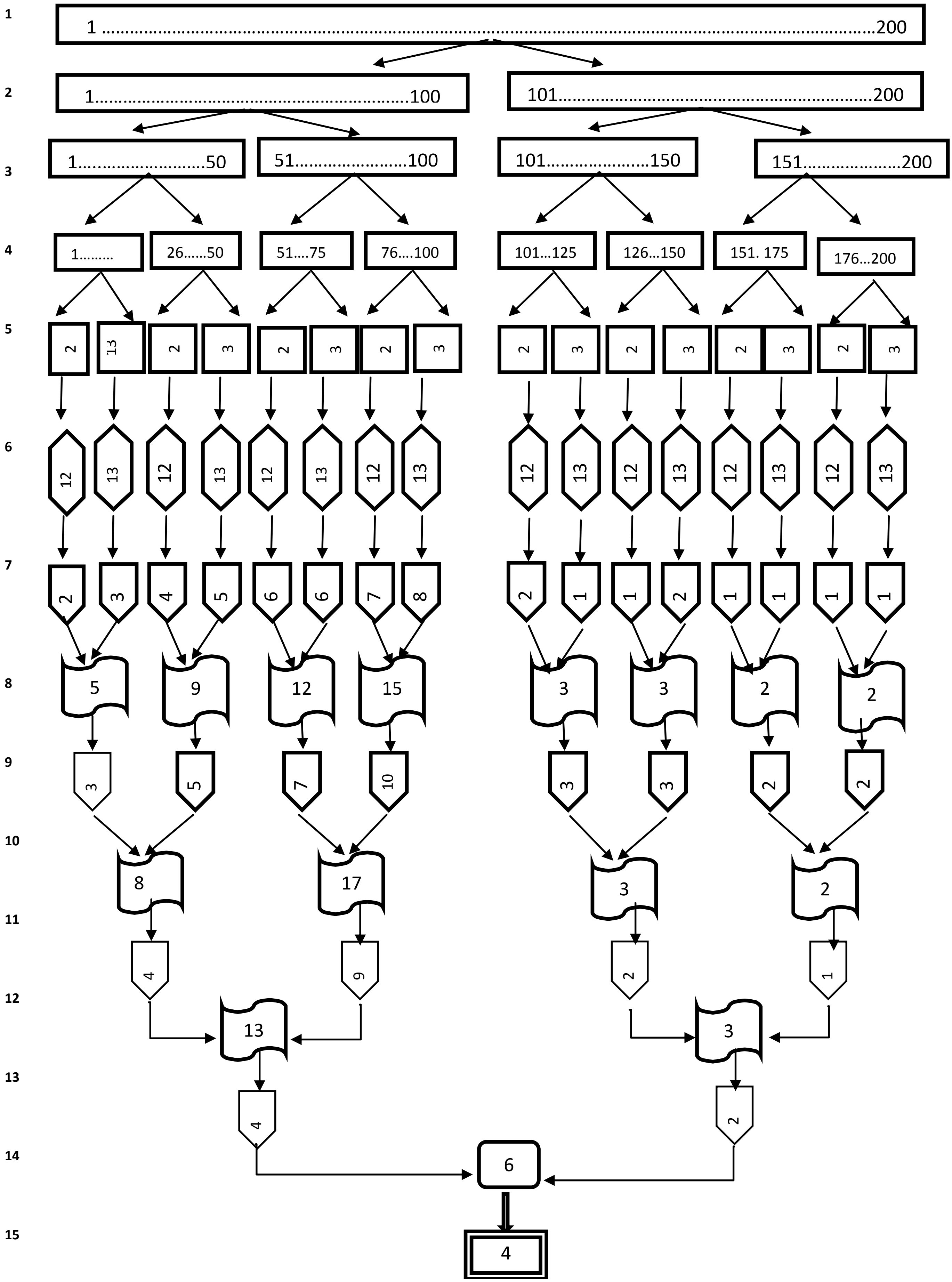


Figure 20: Divide and Conquer using Feature Transformation

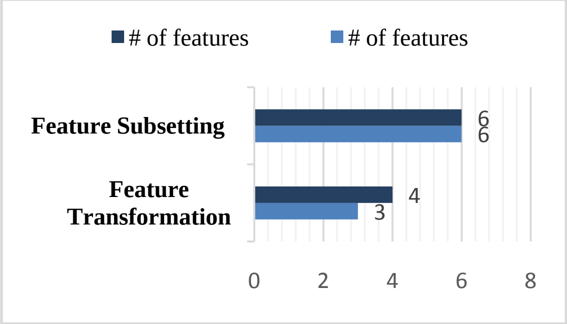


Fig 21: Comparison of the number of attributes necessary for Divide and Conquer Vs Incremental dimensionality reduction

4.3. Comparative analysis between IDR method and Divide and Conquer Paradigm

DnC using FS and FT techniques have yielded the same results when subjected to IDR i.e. 100% classification accuracy with FS and 99.7 % classification accuracy with FT. The results are the same when traditional DR is applied assuming all features are available together.

FS and FT techniques are highly comparable. FS requires lesser attributes than FT to hold the knowledge together (i.e. 1128 Vs 1410). In the AVIRIS mini dataset considered, FS eliminates features which absolutely doesn't add any information or knowledge to the cumulatively accumulated knowledge. FS achieves max classification accuracy by using 6 attributes. While DnC is performed using FT, one stonetower pixel is misclassified as a wheat pixel consistently due to overlapping spectral value and hence achieves a classification accuracy of 99.7% (see Fig 24). The features obtained by DnC are compared with the features obtained by IDR methods for both FS and FT techniques (see Figure 21). Although the number of features match , the features in the reduced space might be slightly different. The features obtained with FS technique for both DnC and IDR method have an average correlation value of 0.87 and the features of FT technique have an average correlation value of 0.99. This indicates that although the feature sets are not exactly the same, the features are highly correlated to each other in both the methods. In DnC method, the procedure breaks down the problem into subproblems, merges the solutions until the problem cannot be reduced any further, thus achieving parallelism. There are two main drawbacks of DnC when compared to IDR method. The first one being - DnC method avails the decision making parameters to the user only at the end of the procedure. The second drawback is that DnC can be applied only when all features are available together not otherwise.

In the proposed IDR method, the procedure can break down the problem(when all features are available) into subproblems , can also factor the problem as a subproblem as it features are made available in the feature space (streaming data) and can build the solution incrementally with whatever data is available at that instant. The solution of the incoming subproblem is incrementally merged with the accumulated solution of the subproblems merged prior , thus building the cumulative knowledge progressively. At any point in time , the Proposed IDR procedure provides the best possible decision making parameters to the user that will be in line with the final, unseen solution that will be achieved at the end. Hence, the user does not have to wait till the end of process to gather the best possible decision making criteria.

5. PERFORMANCE EVALUTAION

How feature transformation and subset selection are targeted depends on the purpose, i.e. whether it is for concept description or for classification. The former aims at preserving the topological structure of the data whereas the latter aims at enhancing the predictive power [23]. A similar rationale applies to FSIDR and FTIDR as well. In a traditional DR approach, one might want to apply DR when there are 200 attributes in the feature space. In Incremental DR, the feature space is regularly and incrementally reduced to compact the feature space and does not wait for features to accumulate before it can reduce. In traditional FS, the maximum number of features needed to achieve maximum classification accuracy is 200 but if an incremental approach is adopted, at the end of IDR, only 14 to 22 attributes will be available in the feature space to hold the same knowledge that will be accumulated by the 200 features. In general, IDR requires approximately 90% less attributes to hold the same knowledge when compared to traditional DR (see Figure 22 and 23).

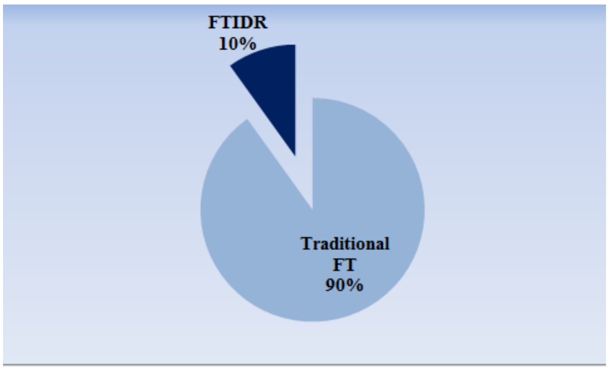


Fig 22: Maximum attributes necessary to hold maximum knowledge using FT method

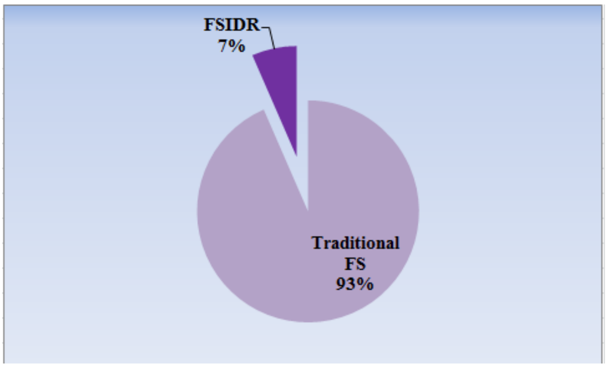


Fig 23: Maximum attributes necessary to hold maximum knowledge using FS method

In the distinct dataset considered, FSIDR achieves 100% classification accuracy and FTIDR achieves 99.7 % classification accuracy. A single pixel in the stone tower class gets wrongly transformed into the spectral band of the alfalfa class thus lowering the accuracy to 99.7%. Both the models are highly comparable with a negligible difference (see Figure 24).

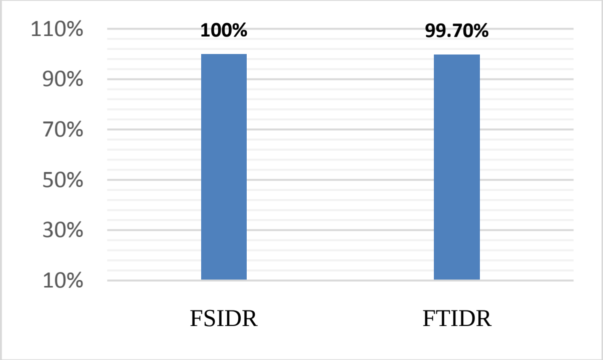


Fig 24: Classification accuracy (%) achieved: FSIDR Vs FTIDR for the hyperspectral - Indian Pines data

Nonetheless, the traditional DR also achieves 99.7% classification accuracy indicating that the proposed approach achieves at least what a traditional DR can achieve. This further supports the IDR model to be perfectly in sync with the traditional DR model. The IDR model scores over the traditional DR as it allows decision making in the middle of streaming data, requires lesser attributes to capture the same knowledge. Hence, the proposed approach is both functionally and computationally more effective.

6. CONCLUSION

Given that data is being generated at a faster pace than ever, the necessity to reduce the features regularly in order to maintain the system performance and to extract useful knowledge for decision making is becoming more evident. In such a scenario, an incremental approach to reduce the feature space, to accumulate the knowledge without looking back at the previous data, will be a remarkable one. In domains where massive feature space builds up in volume within no time as in hyperspectral images, the proposed IDR approach will not only aid in compacting the knowledge globally but also provides a local and optimal knowledge to make decisions in the interim as necessary. The local optimal knowledge will be in line with the unseen data and will be built upon the compacted knowledge accumulated thus far, completely eliminating the need to look back at the previous data. Hence the collection of data space and generation of knowledge from the features gathered at a given point can function in parallel with IDR.

IDR approach can be further applied with association rules to incrementally build the association of features .The association rules obtained thus will aid to statistically measure and identify methods to discover patterns in the feature space. The knowledge thus obtained will aid to determine the implied features, enhance decision making capabilities and identify critical dependencies for risk management and mitigation.

7. REFERENCES

[1] https://purr.purdue.edu/publications/1947/1

[2] http://www.cse.msu.edu/~cse802/Feature_selection.pdf

[3] Plaza, J., Plaza, A. J., & Barra, C. (2009). Multi-Channel Morphological Profiles for Classification of Hyperspectral Images Using Support Vector Machines. Sensors, 9, 196-218.

[4] Zhang, Y., Du, B., Zhang, L., & Liu, T. (2017). Joint Sparse Representation and Multitask Learning for Hyperspectral Target Detection. IEEE Transactions on Geoscience and Remote Sensing, 55(2), 894-906.

[5] http://www.microimages.com/documentation/Tutorials/hyprspec.pdf

[6] http://lesun.weebly.com/hyperspectral-data-set.html AND https://purr.purdue.edu/publications/1947/1

[7] Aggarwal, C. C., & Yu, P. S. (2001, May). Outlier detection for high dimensional data. In ACM Sigmod Record (Vol. 30, No. 2, pp. 37-46). ACM.(sparsity)

[8] Signal Theory Methods in Multispectral Remote Sensing. David A Landgrebe. ISBN: 978-0-471-42028-6. 528 pages. January 2003

[9] Subramanian, S., Gat, N., Ratcliff, A., & Eismann, M. (2000). Real-time hyperspectral data compression using principal components transformation. In In Proceedings of the AVIRIS Earth Science & Applications Workshop.

[10] C.-I Chang, Hyperspectral Data Exploitation: Theory and Applications. New Jersey: John Wiley and Sons, 2007.

[11] P. Zhong, P. Zhang, and R. Wang, “Dynamic learning of SMLR for feature selection and classification of hyperspectral data,” IEEE Geoscience and Remote Sensing Letters, vol. 5, no. 2, pp. 280-284, April 2008

[12] Preet, P., & Batra, S. S. (2015). Feature Selection for classification of hyperspectral data by minimizing a tight bound on the VC dimension. arXiv preprint arXiv:1509.08112.

[13] Pal, Mahesh, and Giles M. Foody. ”Feature selection for classification of hyperspectral data by SVM.” Geoscience and Remote Sensing, IEEE Transactions on 48.5 (2010): 2297-2307.

[14] Agarwal, Abhishek, Tarek El-Ghazawi, Hesham El-Askary, and Jacquline Le-Moigne. ”Efficient hierarchical-PCA dimension reduction for hyperspectral imagery.” In Signal Processing and Information Technology, 2007 IEEE International Symposium on, pp. 353-356. IEEE, 2007.

[15] Syed Zakir Ali., P Nagabhushan., Pradeep Kumar R, Incremental datamining using Clustering Intelligent Methods of Fusing the Knowledge During Incremental Learning via Clustering in A Distributed Environment , PhD Thesis, 2010

[16] Meenakshi, H. N., & Nagabushan, P (2017). Target Class Guided Compression in Feature Subspace, IJCST, 4(6).

[17] Nagabhushan, P., & Mahadev, P. (2014). Incremental Feature Subsetting useful for Big Feature Space Problems. International Journal of Computer Applications, 97(12).

[18] Preeti Mahadev and P Nagabhushan. Incremental Feature Transformation for Temporal Space. International Journal of Computer Applications 145(8):28-38, July 2016

[19] P. Nagabhushan, An efficient method for classifying remotely sensed data (incorporating dimensionality reduction), Ph.D thesis, Universityof Mysore, 1988

[20] Datta, Aloke, Susmita Ghosh, and Ashish Ghosh. "Unsupervised band extraction for hyperspectral images using clustering and kernel principal component

analysis." International Journal of Remote Sensing 38.3 (2017): 850-873.

[21] NASA JPL, AVIRIS Data Portal [online]. Available at http://aviris.jpl.nasa.gov/alt_locator/. [Accessed June 2016].

[22] Principal Component. Analysis, Second Edition. I.T. Joliffe. Springer, NewYork, 2002

[23] Liu, H., & Motoda, H. (1998). Feature transformation and subset selection. IEEE Intell Syst Their Appl, 13(2), 26-28.

[24] https://people.eecs.berkeley.edu/~vazirani/algorithms/chap2.pdf