International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Performance Enhancement of Dimension Reduction for Microarray Data

Shubhangi N. Katole[1], Swapnili P. Karmore[2]

[1]Department of Computer Science and Engineering, G. H. Raisoni College of Engineering, Nagpur, India

[2] Department of Computer Science and Engineering, G. H. Raisoni College of Engineering, Nagpur, India

Abstract: ***Due to the importance of gene expression data in cancer diagnosis and treatment, microarray gene expression data have concerned more and more attentions from cancer researchers in recent years. This paper proposes and implements collision of the employ of dimension reduction methods for the microarray datasets. However, in real-world computational analysis, such data common congregate with the curse of dimensionality due to the tens of thousands of measures of data. Therefore, developing effective dimension reduction method is a tricky problem for high dimensional dataset. Here, we used two Algorithms that is Total PLS and MINE method for dimensional reduction of the microarray data. Next to this step the Hybrid algorithm is applied to the data where in Hybrid algorithm the Total PLS and MINE algorithm are combined. Overall results shown that using Hybrid algorithm provides an improvement in performance of redundancy, efficiency, Accuracy and deviation rate as compared to previous algorithms used.***

Keywords: microarray data, dimension reduction, redundancy, efficiency, MINE, hybrid algorithm

1. Introduction

Dimensionality reduction of the large data becomes important and is full of rising challenges. Dimensionality reduction is a significant task in bioinformatics. The rationale of dimensionality reduction is to reduce, understand and visualize the structure of complex data sets. Dimensionality reduction is the fundamental task for many large-scale information processing problems [1]. Constructing effective dimension reduction methods is decisive. There are several methods for the dimension reduction of the data accessible which are rising but no one is the best method of data dimension reduction for all circumstances for the reason that at the time of the process some information is forever lost [2]. To improve the performance of the data we are going to use the new dimension reduction method. Feature extraction and Feature selection is a universal method of dimension reduction. In Feature extraction method the original high-dimensional feature space is estimated on to low-dimensional feature space. In Feature selection only uses a division of features from the original data [3]. The benefit of feature selection is that the rotation of selected data table does not take place. Therefore, it is simple to interpret the results. The shortcoming is that this method might cause information failure. Information failure does not happen yet, as there are surplus or unrelated features [1]. For typical microarray data analysis, the training sample size is always limited. Due to many data mining classification algorithms may be short of efficiency or even fail in high dimensional microarray data analysis. Dimension reduction is a good choice to variable selection in order to overcome the dimensionality problem. Dimension reduction uses a little quantity of features to substitute a feature subset containing well-built correlations in the original data [4].

How to pick up the accuracy of text classification is an important and hard problem. Given the ever increasing amount of large, high-dimensional data sets acquired in a variety of scientific disciplines and application domains, efficient methods for dimension reduction and feature selection play an essential role in modern data processing. Microarray data has properties of high clamour, high inconsistency, high dimensionality and high correlations [5]. A typical microarray data set would have thousands of variables (genes) and a very small number of (biological) replicates, so regression analysis is difficult in practice. The trouble of dimension reduction is introduced as a approach to defeat the curse of the dimensionality when dealing with vector data in high-dimensional spaces and as a modelling tool for such data. The development of high throughput technology creates large volumes of high-dimensional gene expression data, which is easily available and accessible for data mining community and cancer researchers.

Currently, the significance of gene expression data in cancer analysis and conduct has attracted more and more courtesy from the researchers. In this context, gene expression profiles can categorize dissimilar cancer types or subtypes. Microarray data sets have the characteristics of high dimension and small samples. A typical microarray gene expression dataset contains the expression values of tens of thousands of genes in different samples and commonly, only several tens of genes have unique expression patterns to each cluster of samples [6][15][16], which are called the relevant genes. In contrast, the irrelevant genes have very less help in identifying the cluster members and will lead to two samples with low similarity measured by a similarity function that considers the expression values of all genes in the same cluster [3][17]. Due to a large number of genes being irrelevant to each cluster, it is infeasible for using traditional clustering algorithms to detect the clusters. Therefore, high dimensional data put demands on the efficiency and effectiveness of the learning algorithm. One of the major challenges of microarray data analysis is the overwhelming number of measures of gene expression levels compared with the small number of cancer samples, specifically, the samples displaying different behaviours in only a few of the genes.

1.1 Dimension Reduction

Dimension reduction is an essential step in high dimensional data analysis. It extracts a small number of features by removing irrelevant, redundant, and noisy information. Dimension reduction is a crucial step for the analysis of high-dimensional data. The problem of dimension reduction can be stated as follows: given the p-dimensional random variable $x = (x_1,\ldots,x_p)T$, discover a lower dimensional illustration of it, $s = (s_1,\ldots s_k)T$ through $k << p$, that captures the contented in the novel data according to some condition. The components of s are occasionally called the unseen components. Different fields use different names for the p multivariate vectors, for example, the term "variable" is mostly used in statistics, while "feature" and "attribute" are alternatives commonly used in machine learning [2]. In essence, the goal of dimensionality reduction methods is to map observations, initially represented as high dimensional vectors, into a lower dimension space [4]. The residue of this paper is described as follows. Related work is provided in section II. Section III provides the dataset for the dimension reduction. In section IV proposed work of the project is provided. In section V and VI the proposed work and results for the project is described respectively and the conclusions is provided in Section VII.

2. Related Work

Author W.H.Yang, D.Q.Dai, and H. Yan elucidated at what time and why the dimensionality reduction through SVD works used for pattern classification errands Also, they have planned two methods of dimension reduction. Then offered the innovative uncorrelated discriminant analysis (UDA) algorithm which is helpful to overcoming the inadequacy of the classical LDA. UDA method has good perceptive power, in clarification and appearance variations for face recognition, gene expression data sets. These entail that UDA is an useful and steady linear prejudice approach for high-dimensional data. But they have not developed data precise techniques to covenant with nonlinear phenomena, such as high noise with gene expression data [8].

Author Jian J. Dai, Linh Lieu, and David Rocke deliberate the assessment of three dimension reduction method, i.e. PLS, SIR and PCA also appraised the comparative performance of classification procedures integrating those methods similar to PCA, PLS reduce the complexity of microarray data analysis. PLS method is computationally very proficient. They establish that PLS and SIR was both expensive in dimension reduction and they were extra effective than PCA. The PLS and SIR pedestal classification procedures execute constantly superior than the PCA based method in guess precision. The results are reliable with the scrutiny of the method. The span of the study is still quite limited. The PCA method has restrictions as datasets are extremely non-linear [9].

Author J Hua, W. D.Tembeb, E.R. Dougherty proposed feature-label distribution models. They have been done the comparison between the classification recital of the feature selection methods while training sample sizes are restricted and quantity features are outsized. They have performed different feature selection algorithms for the experiment of mock data and real data. When this is done, it can observe trends in the behaviors of feature-selection algorithms relative to specific model conditions, such as sample size and the numbers of global and heterogeneous markers. The classification inaccuracy and quantity of discovered useful features are computed. Moreover, the models can be easily comprehensive to multi-class distributions, till it has not done. These issues are come out [4].

Author I. Guyon, J. Weston, S. Barnhill, and V. Vapnik proposed and applied the SVM method of Recursive Feature Elimination (RFE) to gene selection. There is trouble i.e. high dimensional data. The experimentation done on two different cancer databases that taking into account mutual information between genes in the gene selection process impacts classification concert. The experimental demonstration is better in classification performance. Confirmation of the biological significance of the genes found by SVMs. The RFE method was demonstrated for linear classifiers, including SVMs. The explorations specify that RFE is more vigorous method to data over fitting than other methods, together with combinatorial search. The inadequacy of SVM-RFE method is elevated computational charge [10].

Author Jun Yan, Benyu Zhang, Ning Liu proposed two dimension reduction algorithms based on OC algorithms and are named as IOC and OCFS algorithms. These methods are helpful for providing best solution according to OC criteria and planned under the equivalent optimization criterion. These two algorithms are good in their performance if the comparison is done with other methods. For data preprocessing of high-dimensional data and to stream data classification problems, they have discussed the relationship between FE and FS approaches under a integrated structure which could help the readers decide other suitable dimensionality reduction algorithms for classification tasks. From the planned structure for dimensionality reduction, one can see that dimensionality reduction could be treating as the reduction of some objective function. IOC is the FE approach which aims reduction in a incessant solution space and OCFS is the FS approach which aims reduction in a distinct solution space [11].

Author Chih-Wei Hsu and Chih-Jen Lin discussed decay implementations for two altogether methods and compared them with three methods based on numerous binary classifiers: one-against-one, one-against-all and DAG. The experiments on great troubles exemplify that one-against-one method and DAG may be more appropriate for realistic exercise. The issue to analysis data among a very outsized amount of classes, that there may have extra differences among these methods if the data set has a small amount of points in many classes [12].

Author Sampath Deegalla, Henrik Bostrom, introduced the innovative dimension reduction method in which combination of two approaches occurred. Feature fusion and classifier were the methods measured in accumulation with the k-NN classifier and two approaches to combine the

results of combination. The experimentation with eight microarray data sets exposed that dimensionality reduction certainly is useful for nearest neighbor classification and that combine the output of these methods which can more pick up the classification accuracy evaluate to the entity dimensionality reduction method. Yet, if one contrasts the individual methods and fusion methods then the feature fusion method achieve the most excellent classification accurateness in the majority all. Hence, to prefer some of the fusion approaches should be favored to select a few of the one dimensionality reduction methods, as the earlier can be probable to direct to accurateness in classification and strength according to the option of number of dimensions. Number of issues is there which will require further study [13].

3. Datasets used in the Experiment

In the proposed work, the Cancer Microarray Datasets is used as input datasets. We select eight types of cancer datasets are measured where high dimensional datasets are normal. We have applied the algorithm the different cancer datasets. It contains DNA microarray gene expression data. The selected microarray datasets have elevated size and numerous types.

4. Proposed Work

Total PLS algorithm is the integrated framework of the feature selection and the feature extraction algorithm. In general, problem is that dimension reduction methods can be classified into linear and nonlinear methods. Usually, it is able to improve the recognition rate through selecting the appropriate kernel function and its parameters.

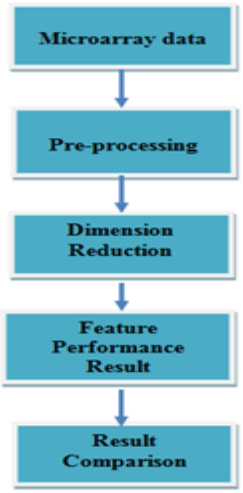

Figure 1: Block Diagram of Proposed Work

However, the resulting model and the results of the data cannot be easily interpreted. So the maximal information-based nonparametric exploration is used to discover the nonlinear relationship among specific genes in microarray data. Also to improve the overall performance we plan to introduce the Hybrid Algorithm in which Total PLS and MINE Algorithm will be combined.

The main objective of this method dimension reduction is to enhance accuracy and improve efficiency of database. To discover the strong relationship among gene coexpression and coregulation, to interpret and understand the results easily are the targets to make the comparison. The purpose of dimension reduction is to find some form of structure hiding in the original large size data. The original large size feature space is anticipated on to low-dimensional feature space. Also occurs improvement in the reduction of data retrieval time and redundancy. The proposed plan of work shown in figure 1

5. Methods and Algorithms

5.1. Clustering Method

In the pre-processing step the k-means clustering method is applied on the data. In k-means clustering method, tries to find a user specific number of clusters (k), which are represented by their centriod, by reducing the square error function. K-means is straightforward and can be used for a large variety of data types. As shown in fig. 2, it gives the steps for K- means clustering.

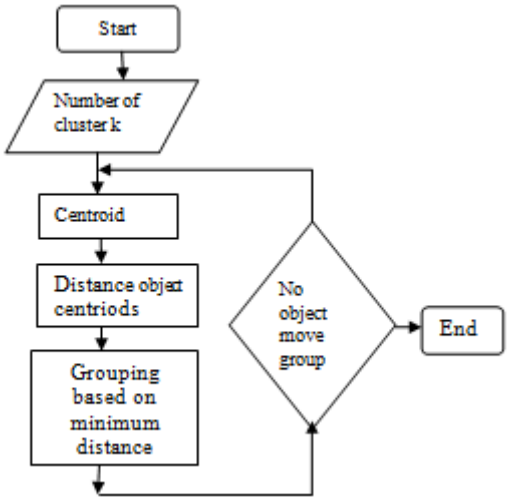

Figure 2: K-means clustering method

5.2. Partial least squares (PLS)

The X-scores and Y-scores are chosen so that the association among successive join up of scores is as tough as possible. In rule, this is like a vigorous form of redundancy analysis, in search of directions in the factor space that are linked with high deviation into the reactions other than prejudicing them to orders that are precisely guessed. Another approach is to put up the PLS model for a specified amount of features lying on single place of data and afterward to check it on another, prefering the number of removed features used for which the entire calculation inaccuracy is reduced.

Another inference method for partial least squares regression components is the SIMPLS algorithm which can be described as follows.

For each h=1,…,c, where A0=X'Y, M0=X'X, C0=I, and c given, compute q_h, the dominant eigenvector of $A_h'A_h$

1. $w_h=A_hq_h$, $c_h=w_h'M_hw_h$, $w_h=w_h/sqrt(c_h)$, and store w_h into W as a column
2. $p_h=M_hw_h$, and store p_h into P as a column
3. $q_h=A_h'w_h$, and store q_h into Q as a column
4. $v_h=C_hp_h$, and $v_h=v_h/\|v_h\|$
5. $C_h+1=C_h - v_hv_h'$ and $M_h+1=Mh - p_hp_h'$
6. $A_h+1=C_hA_h$

The T of SIMPLS is computed as T=XW and B for the regression of Y on X is computed as B=WQ'[1].

5.2.1 Total PLS Algorithm

Total PLS algorithm achieves both PLS-based feature selection and feature extraction in a unified PLS framework, is called as Total PLS dimension reduction Algorithm. Steps for the Total PLS Algorithm are as follows.

Input: Train Xn×p, ClsYn×1, Dim

Output : XScore ‖ XScore is the score on PLS-based latent factor

(1) Initialization
-Encode class label ClsYn×1 and generate Class ClsYn×g
-Set nfac = unique(ClsY) ‖ unique(ClsY) indicates the number of category

(2) Feature Selection
-Obtain idx
PLSRFE(TrainX,ClasssY,nfac)
-Update TrainX, whose features only include top Dim features in idx

(3) Feature Extraction
-For j=1 to nfac do
-Calculate score matrix Tj=<TrainX,wj>
-Update Xscore so that Xscore = [Xscore,Tj]
Return Xscore[1]

5.3. Hybrid Algorithm:

In hybrid algorithm Total PLS Algorithm and MINE Algorithm is combined. We have applied the hybrid algorithm to improve the performance of the dimension reduction of the cancer microarray data. The Hybrid algorithm is applied on the microarray data to improve the performance of the dimension reduction method. Steps for the Total PLS algorithm as above and MINE algorithm are as follows.

5.3.1 Maximal Information- Based Nonparametric Exploration (Mine) Method:

The MINE method is used in statistics. In the MINE Method MIC (Maximal Information Coefficient) plays an important role. MIC is the larger part of the MINE technique. The idea of the MIC is that if the relationship is existing in between the two variable data.

The steps for the MIC are as:

1. Estimating the range of rows and columns i.e. x and y respectively for grid resolution plot two variable data.
2. Partitioning the data. In this step, Grid can be drawn on the disperse plot of the two variables that partitions the data for sum up that relationship of the data.
3. Placements the partitions. In this step the partitions of the data will be placed after grid resolution of the data.
4. Calculating the MIC (Maximal Information Coefficient). Here, to work out the MIC of a set of data which is two variable data, we determine all grids up to a maximal grid declaration, reliant on the sample dimension of the data.
5. Finding the highest mutual information to store each resolution. Reliant on the sample dimension, work out for every pair off of integers (x,y) the major probable common in sequence attainable by any x-by-y grid applied to the data. We name the attribute matrix M = (mx,y), where mx,y is in which the will be occurred the highest possible common information.
6. Normalization. We then normalize these common information values to make certain a fair comparison between grids of different dimensions and to get modified values among 0 and 1.
7. Storing normalized mutual information in the characteristic matrics M(x,y). We define the characteristic matrix M = (mx,y), where mx,y is the highest normalized common information attained by any x-by-y grid, and the statistic MIC to be the maximum value in M [3].

6. Results

In the experiments, we have examined the performance of Total PLS (partial least algorithm), Maximal Information-based Non parametric Exploration (MINE) Method. As a result, Hybrid algorithm is the combination of both algorithms (Total PLS and Maximal Information-based Non parametric Exploration (MINE) Method) that is Hybrid algorithm is drastically improved against those of Total PLS algorithm. We present a technique that improves performance of microarray data dimension reduction. We studied and implement dimension reduction algorithm for enhancing performance of dimension reduction.

Figure 3 shows the redundancy of the data by using Hybrid Algorithm which has the improved performance than previous algorithm. In this figure the non repeated and repeated data are shown.

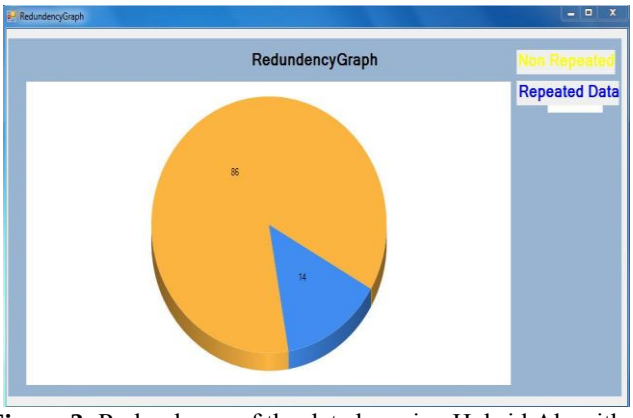

Figure 3: Redundancy of the data by using Hybrid Algorithm

Figure 4 shows the recognition accuracy by using hybrid algorithm is more than total PLS. It indicates improvement in the result. As shown in Figure 5, the results from TotalPLS and Hybrid algorithm shown in terms of the standard deviation which indicates the difference in the performance of both method. Hybrid algorithm can pick up the prediction accuracy and is more constant.

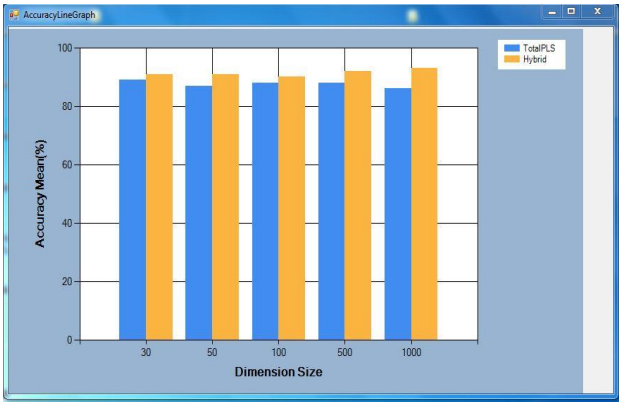

Figure 4: Improved Recognition rate mean accuracy by Hybrid algorithm

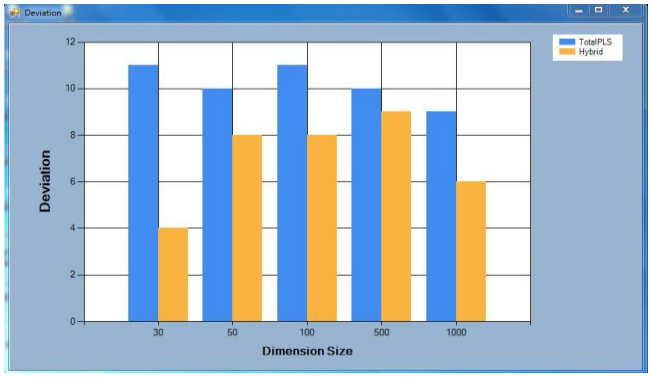

Figure 5: Perfomance analysis of Standard deviation based on Total PLS and Hybrid algorithm

The relationship between the number of chosen feature and recognition accuracy on test sets are recognized in Figure 6. The recognition rate from Hybrid method is much better than the TotalPLS methods.

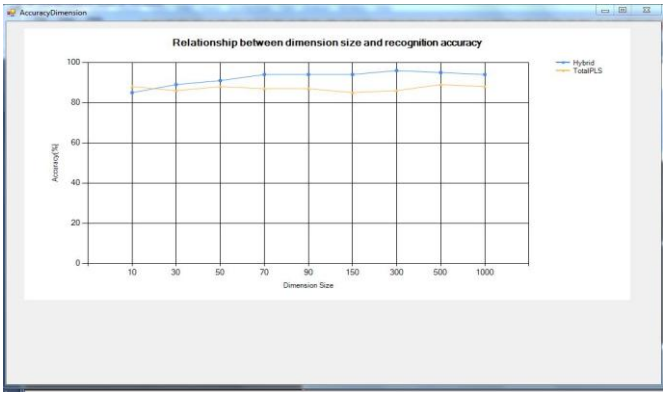

Figure 6: Performance analysis of both algorithm which shows relationship between dimension size and recognition accuracy

Figure 7 shows the efficiency of the Hybrid algorithm is more than the Total PLS algorithm that means the computing speed of the hybrid algorithm is good.

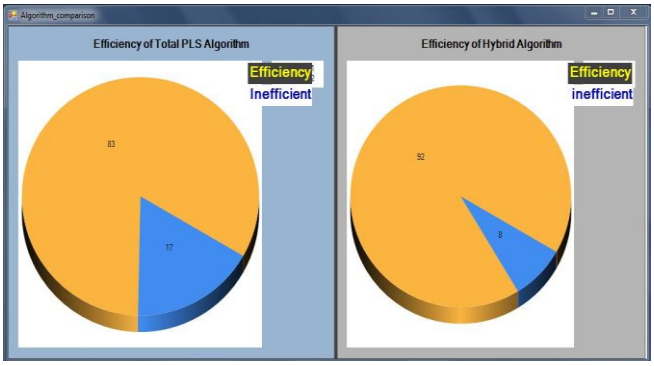

Figure 7: Comparision in Efficiency of the Total PLS and Hybrid algorithm

7. Conclusion

In this paper, we studied and implement dimension reduction algorithm for enhancing performance of the data dimension reduction. In proposed method used Total PLS and MINE algorithm which are combined as in the form of Hybrid algorithm. Using proposed method, we can minimize the time, redundancy of data analysis. We need an important goal of data mining having reasonable and effective access to useful knowledge. The aim of the study was to enhance performs data dimension reduction for the microarray data analysis. This paper introduced the two algorithms. The proposed algorithm that is Hybrid algorithm gives the improved performance in terms of efficiency, recognition accuracy, redundancy of the data as compared to the previous algorithm.

References

[1] Wenjie You, Z Yang, M Yuan, and Guoli Ji "Total PLS: Local Dimension Reduction for Multicategory Microarray Data," IEEE Trans. on human-machine systems, vol. 44, no. 1, february 2014.
[2] D. Araujo and A .D. Neto and A. Martins and J. Melo, "Comparative Study on Dimension Reduction Techniques for Cluster Analysis of Microarray Data," Proceedings of International Joint Conference on Neural Networks, San Jose, California, USA, July 31 – August 5, 2011.
[3] I. Gheyas and L. Smith, "Feature subset selection in large dimensionality domains," Pattern Recognit., vol. 43, no. 1, pp. 5–13, 2010.
[4] J. Hua, W. D. Tembe, and E. R. Doughertya, "Performance of feature selection methods in the classification of high-dimension data," Pattern Recognit., vol. 42, no. 3, pp. 409– 424, 2009.
[5] A. Anaissi, P.J. Kenned, M. Goyal "A Framework for High Dimensional Data Reduction in the Microarray Domain," 9781-42446439-5/10/$26.00 ©2010 IEEE.
[6] A. K. Jain, R. P.W. Duin, and J. Mao, "Statistical pattern recognition: A review," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 1, pp. 4–37, Jan. 2000.
[7] J. Fan and Y. Fan, "High-dimensional classification using features annealed independence rules," Ann. Statist., vol. 36, no. 6, pp. 2605– 2637, 2008.
[8] W. H. Yang, D. Q. Dai, and H. Yan, "Feature extraction and uncorrelated discriminant analysis for high-

dimensional data," IEEE Trans. Knowl. Data Eng., vol. 20, no. 5, pp. 601–614, May 2008.
[9] J. J. Dai, L. Lieu, and D. Rocke, "Dimension reduction for classification with gene expression microarray data," Statist. Appl. Genet. Mol. Biol., vol. 5, no. 1, pp. 1–19, 2006
[10] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Mach. Learn., vol. 46, no. 1–3, pp. 389–422, 2002.
[11] J. Yan, B. Zhang, N. Liu, S. Yan, Q. Cheng, W. Fan, Q.Yang, W. Xi, and Z. Chen, "Effective and efficient dimensionality reduction for largescale and streaming data preprocessing," IEEE Trans. Knowl. Data Eng., vol. 18, no. 3, pp. 320–333, Mar. 2006.
[12] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," IEEE Trans. Neural New., vol. 13, no. 2, pp. 415–425, Mar. 2002
[13] S.Deegalla, H. Bostrom, Fusion of Dimensionality Reduction Methods: A Case Study in Microarray Classification" 12th International Conference on Information Fusion Seattle, WA, USA, July 6-9, 2009.
[14] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G.McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," Science, vol. 334, pp. 1518–1524, 2011.
[15] Yu Wanga, Igor V. Tetkoa, Mark A. Hallb, Eibe Frankb, "Gene Selection from microarray data for cancer classification-a machine learning approach", in Proc. Computational Biology and Chemistry, 29 (2005) 37–46.
[16] S. T. Roweis and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," Science, vol. 290, no. 5500, pp.2323–2326, December 2000.
[17] I. Fodor, "A Survey of Dimension Reduction Techniques" Lawrence Livermore National Lab., CA (US) UCRL- ID-148494, 2002.