Survey of Dimensionality Reduction and Mining Techniques on Scientific Data

D. Lakshmi Padmaja
Associate Professor, C.V.S.R.College of Engineering,
Hyderabad, India

Dr.B. Vishnuvardhan
Professor, CSE Dept, JNTUH College of Engineering,
Nachupally, Karimnagar (Dist), India.

Abstract-Dimensionality reduction techniques on scientific data are currently a focused approach to understand the underlying scientific knowledge in a dataset resulted from scientific experiments and simulations. The sole purpose of the survey paper is to provide comprehension of different dimensionality reduction techniques which are used in the field of scientific data mining. Feature extraction and feature selection are the important techniques of dimensionality reduction; the former removes certain features by way of transformation, where as the later reconstructs its features into a lower dimension space without impairing its initial characteristics. This paper presents various techniques used for mining the scientific data.

Keywords: Dimensionality Reduction, Feature Extraction, Feature Selection, Scientific Data.

I. INTRODUCTION

The objective of the data mining is to extract knowledgeable information from large data sets. It is a combination of multiple disciplines and techniques to solve the extremely challenging problems in the field of bio informatics, astrophysics, weather forecasting, cancer classification, astronomy, Eco System, Modelling, Fluid Dynamics, Structural Mechanics etc. [1,2,3]. It is an interdisciplinary science ranging from statistics to information processing, database systems, artificial intelligence and soft computing. The central theme of data mining is to view the data from different angles to understand the hidden message in the data set. Due to the latest improvements in the processing power and steep storage cost reduction, huge volumes of data has been captured from the scientific experiments and simulations.This has resulted curse of dimensionality by challenging the suitability of existing data mining methods to retrieve the scientific knowledge from the collecteddata.

Many data mining algorithms such as k means, failed to scale well when applied on large size datasets.On the other hand, the same methods are widely used in business applications assuming thatthe data is clean and available in the form of database, which is very rare in scientific domain. Historical data is widely used to generate a large training data set for commercial applications, whereas the training sets in scientific data arehighly sparse in nature and mostly in the unbalanced form. The problems encountered in science and engineering domain is very different from business driven application domain [2, 3]. There is a greater focus on pre and post processing steps which are critical and time consuming tasks of mining scientific data. Hence, the data mining methods which are applied on scientific data requires special attention to retrieve the hidden information rather than pure pattern matching which is prevalent in commercial datamining applications.

Mark J. Embrechset al defined [3] scientific data mining as a technique applied to scientific problems rather than data base marketing, finance, or business driven applications. Scientific Datamining distinguishes itself in the sense that the natures of data sets are different from traditional market driven datamining applications. The datasets involved vast amounts of precise and continuous data and accounting for underlying system non-linarites which are extremely challenging from a machine learning point of view. Two categories of techniques are popular in dimensionality reduction firstly original feature are transformed into lower dimensional space (feature extraction) and secondly a subset of original feature set is selected (feature selection).

Principal Component Analysis(PCA), Multi DimensionalScaling(MDS),Independent Component Analysis(ICA), ISOMAP etc. are used in reducing the dimensionality of large data sets, which are called feature extraction techniques. As the complexity of data increases the performance decreases and often results imprecise output. In the recent past large number of feature selection techniques for dimensionality reduction have been proposed to handle complex highdimensional data. Various studies have been shown the feature selection techniques are superior to their feature extraction counter parts [2, 26].

This paper is organised as follows. In section 1 Introduction, in section 2 definition of the dimensionality reduction is explained. In section 3 briefly discusses about the Literature Survey on Feature Extraction techniques for dimensionality reduction, section 4Literature Survey on the Feature subset selection techniques for

dimensionality reduction and scoring algorithms are presented, which are used in dimensionality reduction. Finally conclusions were given in section 5.

II. DIMENSIONALITY REDUCTION

The dimensionality reduction is defined as follows: Consider a large data set, which is an outcome of scientific experiments, represented in a M x N matrix X consisting of N data vectors, xi (i Є {1, 2…M}) with a dimensionality N. This technique transform data set X with dimensionality n into a new data set Y with dimensionality d while retaining the geometry of the data as much as possible.In general it is difficult to understand geometry of the data clearly. Hence dimensionality reduction is addressed by assuming some critical properties of the data. The following figure describes the taxonomy of techniques for dimensionality reduction [18].

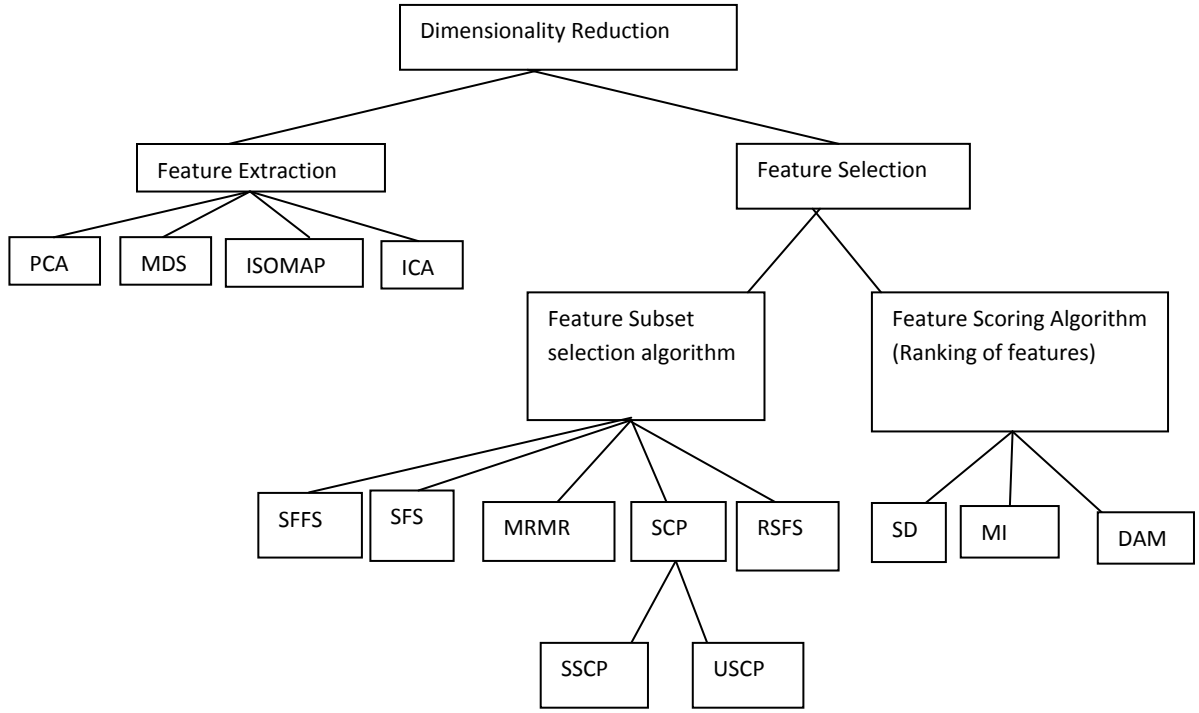

Figure 1: Taxonomy of techniques for dimensionality reduction.

There are two categories of techniques used for dimensionality reduction namely feature extraction and feature selection. The feature extraction techniques assume that the data lie on or near a liner subspace whereas feature selection does not rely on the linearity assumption. Further subdivisions in the taxonomy are discussed in the following sections.

III. LITERATURE SURVEY OF FEATURE EXTRACTION TECHNIQUES FOR DIMENSIONALITY REDUCTION

Feature extraction /transformation are a process of which a new set of features is created. The feature extraction may be a linear or nonlinear combination of original features. There are various techniques to do so PCA is the most popular feature extraction techniques in addition to MDS, ISOMAP, and ICA etc. as mentioned in the Figure 1.

1. Multi-Dimensional Scaling (MDS)

Multidimensional scaling (MDS) is a technique of visualizing the level of similarity of individual cases of a dataset. It refers to a set of related ordination techniques used in information visualization, in particular to display the information contained in a distance matrix. S.H.Bae et al. have described [30] in their paper for solving high dimensional data visualization using MDS as an approach. The solution discussed in theirpaper to handle High Dimensional Data set with efficient use of resources such as CPU and Memory. The MDS algorithm is suitably modified as SMACOF (Scaling by Majoriging A Complicated Function). They proposed a threading based shared memory parallel implementation of SMACOF algorithm to enhance the performance. As the algorithm is realized using Message Passing Interface(MPI) for utilizing distributed memory cluster system, the runtime efficiency decreased.

2. Independent Component Analysis (ICA)

In signal processing, Independent Component Analysis (ICA) is a computational method for separating a multivariate signal into additive subcomponents. This is carried out by assuming that the subcomponents are non-Gaussian signals and that they are statistically independent from each other. ICA is a special case of blind source separation.

J.Basak et al. 2004 in their paper provided [6] techniques for determining the independent components in the Multidimensional data sets and observed that the stable pattern on climate phenomena. This was obtained by Independent Component Analysis (ICA) matched with the physical patterns of oscillations in mean Sea Level Pressure (SLP. The results are verified by finding the linear fit of the independent components provided by the meteorological measurements.

However it is assumed that the independent components are linear in nature. The assumption impacts the accuracy negatively. Hence the authors suggested for nonlinear ICA technique to enhance the accuracy.

Fodor and Chandrika Kamath used [7] ICA to separate the signals in climate data. The ICA technique combined with PCA resulted more promising outcome on dimensionality reduction used on ENSO variation data sets. Currently authors are investigating possible non linearity's in the mixing processes, and providing uncertainty estimates for the estimated components using PCA and ICA.

3. Partial Least Squares (PLS)

Shen and Tan proposed [8] in their paper an innovative method for dimension reduction using penalized logistic regression (PLR) combined with Partial Least Squares (PLS). The method is applied on microarray data for cancer classification and its accuracy is quite competitive when compared to Support Vector Machine (SVM). The overall computational complexity for PLS-PLR is much less than SVM method. They also demonstrated PLR is most robust and can perform very well on the multivariate or covariate matrices.

IV. LITERATURE SURVEY ON THE FEATURE SUBSET SELECTION TECHNIQUES FOR DIMENSIONALITY REDUCTION

Feature selection is the prime the focusof interest quite some time and much work is done in the area of high dimensionality reduction especially using the scientific data sets. With the generation of huge data sets as the outcome of scientific experiments and the requirement for the good machine learning techniques, novel approaches to feature selection are in demand.

1. Feature Subset Selection

The feature selection can be formulated as the problem of finding the best possible subsets S of feature from an initial, very large datasets of features F .where S ⊂ F. Learning of more compact set of features results enhanced classification performance due to removal of unreliable features, lower computational cost and better handson understanding of the classification problem. As the ultimate goal is to enhance the accuracy of classification of data, it is appropriate to define the optimal subset of features to achieve the same. This is achieved by defining a criteria function G(S, D, M) =C. Where D is data set used, M denotes the classification model and S is subset of features of F.The feature subset selection is divided into three divisions namely filter, wrapper and embedded [9, 10, 11]. Q Song et al. [12] proposed a Minimal Spanning Tree (MST) based algorithm for feature subset selection. In first step features are divided into clusters and then the most representative feature is selected from each cluster to form the final subset. The algorithm was tested on 35 data sets such as image, microarray, text data etc. The text data classification accuracy decreased due to FAST implementation. FAST ranks 5 when compared to the other algorithms Ripper, FCFS, CFS etc.

1.1. Sequential Forward Selection (SFS):

SFS proposed [3] by Whitney (1971), the feature set is iteratively updated by including; in each step the feature results in maximal score [13]. Thus the feature set of size d is given by

$$S_d = S_{d-1} \cup argmax_f G(S_{d-1} \cup f, D, M).$$

The function is evaluated with the help of kNN classifier on the subset of data. The value of maximum Unweight Average Recall (UAR) over a range of k values varying from 5 to 150[13]. This measure is the class specific correct classification rate averaged over the existing classes. The SFS iterates upto a maximum of 500 features before selecting the feature set [14, 15].

As $S = argmax_j G\ (S_d, D, M)$

This algorithm is best suited when the subsets are containing small number of features typically less than 9. The basic search starts with an empty set and produces efficient final subset of the highdimensional data. However sometimes it may results suboptimal subset feature selection. It also suffers from high computational cost due to nesting of features of the subsets[16,14].

1.2. Sequential Floating Forward Selection (SFFS)

Inorder to eliminate the problems in SFS Reunanen (2003) proposed [17][15] a simple method called Sequential Floating Forward Selection(SFFS). Though this method has exceptional performance level but often suffers from high computational cost when applied on large scientific datasets. In majority of the cases, this method failed to converge to a stable feature set of definite size and also unable to reach the predefined maximum set size i.e., 500 features in a reasonable computational time[18,19].

1.3. Minimal Redundancy Maximal Relevance (MRMR). Feature subset selection

This approach is proposed by Peng et.al (2005) [27]. The method analyses the mutual information between the features and labels to maximise the relevance while also considering among the features in the selected feature set to minimise redundancy [20, 13, 27]. The algorithm selects the features that are mutually away from each other while maintaining high correlation to the classification variable. This is an approximation method to maximise the dependency between the joint distribution of selected features and variable.

1.4. Set Covering Problem (SCP)

Set Covering Problem (SCP) technique is used for covering all data points using minimum number of features. This is carried out in two steps. First an evaluation set is generated using single feature separately then enough features are selected to complete the correct classifications. The single feature classification is either supervised or unsupervised based learning algorithm using predominantly Gaussian Mixture models to model the single dimensional probability distributions [13]. This method has many applications such as crew scheduling for airlines or railways etc. Depending on whether the method was based on classification using GMMs trained in a completely supervised or a partially unsupervised manner the method is termed as Supervised SCP (SSCP) or Unsupervised SCP (USCP).

1.5. Random Subset Feature Selection (RSFS)

Random Subset Feature Selection (RSFS) technique is used to discover an average feature set which is better than the available one. The features are selected by repetitively selecting from a random subset of features for all possible set of features. Then with the help of classifier like kNN or Bayesian the feature classification is accomplished. In each iteration the relevance of each feature is modified based on the performance of the subset, where the features participate in. After several iterations the feature set quality improves gradually as random components in the selection process become averaged out. In the same way each feature is evaluated in terms of its useful ness when compared to many other feature combinations. As the current values are not dependent on the previous choices, the method is not susceptible to local optimal solution. This method is based on the concept of Random Forest proposed by (Breiman, 2001) [21] and Random kNN (RKNN, Li et al., 2011) [22], where classification is divided into a set of classifiers that uses random subsets of features. The quality of individual feature is evaluated based on its participation in correct classification.

2. Feature Scoring Algorithms

When compared to feature subset selection methods, feature scoring algorithms provide a score value for each feature to understand its usefulness [23,24]. In order to use feature scoring methods for subset determination the size of the subset play critical role in determining the accuracy of the outcome. The widely used feature scoring methods are Statistical Dependency (SD) and Distribution Alignment Matching (DAM)[13].

2.1. Statistical Dependency between features and labels.

The goal of Statistical Dependency method (SD) is to measure weather the values of a feature are dependent on the class labels or not. Each feature value first divided into oneof the Qs levels where the quantization scale is adaptively determined such that each category contains an equal amount of samples over the entire data set [25, 13, 26]. Instead of a conventional uniform division scale, some statistical validity to the occurrence of different division scales are chosen for this purpose. Statistical Dependency between feature values Y and class labels Z is evaluated.

The larger the SD the higher is the dependency between the feature value and class labels. If the feature is fully independent then SD attain the minimum value of 1.However the SD measure is more sensitive to individual division levels due to the absence of compression. By using Mutual Information (MI) it is to possible to reduce the sensitivity to individual division levels [26, 24, 27].

Both SD and MI methods are used for scoring and ranking of features the chosen number of features having the highest values can be selected [28] for further processing.

2.2. Distribution Alignment and Matching (DAM):

So far the methods that are used in feature selection are based on labelled data. This is also known as an offline approach to select the features. The Distribution Alignment and Matching is a feature scoring method and is fully unsupervised as it does not make use of class labels [13]. Though DAM, it results less accurate output, but complement the conventional algorithms.

V. CONCLUSION

The paper presents a comprehensive survey of different dimensionality reduction techniques in the field of scientific data mining. From the survey it is clear that feature subset selection methods are superior when compared to feature extraction techniques. In future it is possible to enhance the performance of the previously mentioned techniques under subset feature selection category. In addition, a shift in focus towards the development of subset feature selection techniques is useful for mining the complex scientific data sets.

REFERENCES

[1] Usama Fayyad, David Haussler, and Paul Stolorz,Mining Scientific Data, ACM November 1996.
[2] Chandrika Kamath, Scientific Data Mining A PRACTICAL PERSPECTIVE ,Society for Industrial and Applied Mathematics(SIAM) 2009.
[3] Mark J. Embrechts, Boleslaw Szymanski, KarstenSternickel ,Introduction to Scientific Data Mining: Direct Kernel Methods & Applications : Chapter 10,Computationally Intelligent Hybrid Systems : The Fusion of Soft Computing and Hard Computing, Wiley, New York, pp 317-365.
[4] Robert L. Grossman, Chandrika Kamath ,Philip Kegelmeyer, Vipin Kumar and Raju R. Namburu, Data Mining for Scientific and Engineering Applications, Springer Science + Business Media Dordrecht, Kluwer Academic Publishers in 2001.
[5] Whitney, A.W.,A direct method of nonparametric measurement selection, IEEE transactions on Computers 20, 1100-1103, 1971.
[6] JayantaBasak,AnantSudarshan, Deepak Trivedi and M.S.Santhanam,Weather Data mining Using Independent Component Analysis, Journal of Machine Learning Research5(2004)239-253.
[7] Imola K. Fodor and Chandrika Kamath, Using Independent Component Analysis to separate signals in climate data, Lawrence Livermore National Laboratory,CA,USA.
[8] Shen and Tan ,Dimension Reduction Based Penalized Logistic Regression for Cancer Classification using Microarray Data, 2005.
[9] J. Han and M. Kamber. Data Mining: Concepts and Techniques, Morgan Kaufman, 2001.
[10] RonKohaviand GeorgeH.John, Wrappersforfeaturesubset selection, 97(1):273-324.
[11] SanmayDas, Filters,wrappersandaboosting-basedhybridforfeatureselection
[12] J.N.Qinbao Song, G.Wang, A fast clustering-based feature subset selection algorithm for high dimensional data: IEEE Transactions on Knowledge and Data Engineering, Vol.25, IEEE.
[13] JouniPohjalainen,OkkoRasanenandSerdarKadioglu, Feature Selection methods and their combinations in high dimensional classification of speaker likability, intelligibility and personality traits.
[14] SatryaFajriPratama,AzahKamilahMuda,Yun-HuoyChooandNoorAzilahMuda, Computationallyinexpensivesequentialforwardfloating selection for acquiringsignificantfeaturesforauthorshipinvariancenessinwriteridentification.1(3):581–593.
[15] Pudil, P., Novovicova. J.,Kittler.J., Floating search methods in feature selection, Pattern Recognition Letters 15, 1119-1125, Elsevier,1994.
[16] H. Liu and H. Motoda. Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic Publishers, 1998.
[17] Reunanen, Overfitting in making comparisions between variable selection methods, Journal of Machine Learning Research3,1371-1382, Elsevier, 2003
[18] L.J.P.vanderMaaten,E.O.Postman,andH.JvandenHerik, Dimensionalityreduction:Acomparative review, Elsevier 2008.
[19] GokhanGulgezen,ZehraCataltepeandLeiYu, Stableandaccuratefeatureselection.
[20] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. In Proceedings of the Computational Systems Bioinformatics conference (CSB'03), pages 523-529, 2003.
[21] Breiman, L., Random forests. Machine Learning 3, 5–32. Elsevier,2001.
[22] Li, S., Harner, J., Adjeroh, D., Random kNN feature selectiona fast and stable alternative to random forests, BMC Bioinformatics12., Elsevier.,2011.
[23] I. Guyon and A. Elissee .An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157-1182, 2003.
[24] C.Krier, D.Francois, V.Wertz,M.Verleysen, Feature scoring by mutual information for classification of mass spectra, pp. 557-564
[25] Lei Yu, Jessica L Rennert, Huan Liu, and Michael E Berens. Exploiting statistical redundancy in expression microarray data to foster biological relevancy.Technical report, Department of Computer Science and Engineering, Arizona State Univeristy, 2005.TR-05-005.
[26] Zheng Zhao, Huan Liu et al. Advanced Feature Selection research.
[27] F. Ding C. Peng, H. Long. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE TRANSACTIONS ON PATTERN ANAL-YSIS AND MACHINE INTELLIGENCE, 27(8):1226-1238, 2005.
[28] Geng X.,Liu,T. Qin, T.,Li, H, Feature selection for ranking, in: 30th Annual Intl. ACM,ACM Press, p. 407-414.
[29] Veronica-Bolon-Canedo, et al, Statistical Dependence measure for feature selection in microarry datasets, ESANN, April 2011,
[30] Seug- HeeBae, Judy Qiu and Geoffrey Fox, High Performance Multi Dimensional Scaling for Large High-Dimensional Data Visualization, IEEE Transaction of Parallel and Distributed System, Vol 0, No 0, January 2012.
[31] Mark A. Hall and Lloyd A. Smith. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper, 1999.
[32] K. Kira and L.A. Rendell. A practical approach to feature selection. In Sleeman and P. Edwards, editors, Proceedings of the Ninth International Conference on Machine Learning (ICML-92), pages 249-256. Morgan Kaufmann, 1992.
[33] Reunanen, Over fitting in feature selection:Pitfalls and solutions, Aalto university Espoo,Finland, Ph.D,Thesis, 2012.
[34] G.H.John, R.Kohavi and K.Peger. Irrelevant feature and the subset selection problem, Proceedings of the Eleventh International Conference ,pages 121-129.