Dimension diminution of Data objects: A survey

[1] Nootan Agrawal,[2]Mr. Sandeep Gonnade
[1] (M.Tech Student) CSE Department, [2] (Assistant Professor) CSE Department
MATS University, Raipur(C.G.), India

***Abstract* - Data mining application has massive advantages, as historical data have huge number of features. Feature selection is an essential role in improving the eminence of learning algorithms in data mining and machine. This has been broadly deliberated in supervised learning, whereas it is still comparatively infrequent researched in case of unsupervised learning. Each data mining application has familiar matter; dataset has huge number of features which is immaterial or redundant to the data mining job in hand which pessimistically affects the performance of the fundamental learning algorithms, and makes them less efficient. Henceforth reducing the dimensionality of dataset is primary and important job for data mining applications and machine learning algorithms so that computational burden of the learning algorithms can be minimized. In this paper we will discuss different feature selection algorithms so as to find out factors which affect the performance of existing algorithm so that we can move further for researching another novel method for data mining application.**

***IndexTerms* - Component,formatting,style,styling,insert.**

I. INTRODUCTION

Feature selection is an essential role in improving the eminence of learning algorithms in data mining and machine. This has been broadly deliberated in supervised learning, whereas it is still comparatively infrequent researched in case of unsupervised learning. As per Dunham (2002), machine learning tasks can be seen as predictive or descriptive ones. Classification is an example of predictive models. Friedman (1997) described it as a model where discrete output values (class labels) are learnt from the different variables (features) of the input data. Clustering, on the other hand, is categorized by Dunham (2002) as a descriptive task. The features of the input data are used to categorize it without supervised training. In both cases, the choice of the feature-set plays an important role in the performance of the data mining problem. Liu et al. (2010) listed three advantages for removing irrelevant and redundant features: it makes the data mining task more efficient, improves its accuracy and simplifies the inferred model, making it more comprehensible.

By unsupervised learning we stand for unsupervised clustering. Clustering is the procedure of finding groupings by combining "similar" founded on some similarity measure objects collectively. For numerous learning domains, human being defines the features that are potentially functional. However, not all of these features may be relevant. In such a case, choosing a subset of the original features will often lead to better performance.

Feature selection is popular in supervised learning (Fukunaga, 1990; Almuallim & Dietterich, 1991; Cardie, 1993; Kohavi & John, 1997). For supervised learning, feature selection algorithms maximize some function of predictive accuracy. Because we are given class labels, it is natural that we want to maintain only the features that are interrelated to or lead to these classes. But in case of unsupervised learning, class labels not given. Which features should we keep? Why not use all the information we have? The problem is that not all features are important. Some of the features may be redundant, some may be irrelevant, and some can even misguide clustering results. In addition, reducing the number of features increases comprehensibility and ameliorates the problem that some unsupervised learning algorithms break down with high dimensional data.

Concluding that reducing the dimension of data set has following advantages:

- It reduces the storage, time and space required.
- Elimination of multi-co linearity improves the feat of the machine learning algorithm.
- It becomes easier to think about the data when reduced to low dimensions such as 2D or 3D.

The remaining sections of the paper are organized as section II we will converse about previous work has been carried out in this field. Further in section III we will discuss about how we motivated for research work. In section IV we will identify problems in dimension reduction approach. Section V converse the comparison of some existing algorithms. Section VI concludes our survey.

II. LITERATURE SURVEY

Numerous research works has been carried for dimension reduction of data instance to exploit feature.

Ahmed Elgohary, Ali Ghodsi & Ahmed K. Farahat (2013) proposes algorithm that depends on a novel recursive formula for the reconstruction error of the data matrix, which allows a greedy selection criterion to be calculated efficiently at each iteration. They have also presents an accurate and efficient MapReduce algorithm for selecting a subset of columns from a massively distributed matrix. This work enables data analysts to comprehend the insights of the data instance and explore its

secreted structure. The preferred data instances can also be used for data preprocessing tasks such as learning a low-dimensional embedding of the data points.

Ahmed K. Farahat, Ali Ghodsi, and Mohamed S. Kamel (2013) defines a generalized column subset selection problem which is concerned with the selection of a few columns from a source matrix A that best approximate the span of a target matrix B. They proposes a fast greedy algorithm for solving this problem and draws connections to different problems that can be efficiently solved using the proposed algorithm.

Carlos Vicient (2012) discussed about log jam introduced by the manual semantic mapping process. To deal with this problem, presents a domain-independent, automatic and unsupervised method to detect relevant features from heterogeneous textual resources, associating them to concepts modeled in background ontology. The method has been applied to raw text resources and also to semi structured ones (Wikipedia articles). The work has been weathered in the Tourism domain, showing promising results.

Ahmed K. Farahat (2011) presents a novel greedy algorithm for unsupervised feature selection. The algorithm optimizes a feature selection standard which measures the reconstruction error of the data matrix based on the subset of selected features. Ahmed K. Farahat proposes a novel recursive formula for calculating the feature selection criterion, which is then employed to develop an efficient greedy algorithm for feature selection. Additionally two memory and time efficient variants of the feature selection algorithm are proposed.

Yi Yang & Heng Tao Shen (2011) discussed that it is much more complicated to select the discriminative features in unsupervised learning due to be deficient in of label information. They have proposed a new unsupervised feature selection algorithm which is able to select discriminative features in batch mode. An efficient algorithm is proposed to optimize the l2,1-norm regularized minimization problem with orthogonal constraint. Different from existing algorithms which select the features which best preserve data structure of the whole feature set

Jennifer G. Dy & Carla E. Brodley (2003) identified two issues involved in developing an automated feature subset selection algorithm for unlabeled data: the need for finding the number of clusters in conjunction with feature selection, and the need for normalizing the bias of feature selection criteria with respect to dimension. They present proofs on the dimensionality biases of these feature criteria, and present a cross-projection normalization scheme that can be applied to any criterion to ameliorate these biases.

Earlier methods used for dimension reduction:

i) PCA-LRG: is a PCA-based method that selects features associated with the first *k* principal components. It has been shown that by Masaeli et al. that this method achieves a low reconstruction error of the data matrix compared to other PCA-based methods.
ii) FSFS: is the Feature Selection using Feature Similarity method with the maximal information compression as the feature similarity measure.
iii) LS: is the Laplacian Score (LS) method.
iv) SPEC: is the spectral feature selection method using all the eigenvectors of the graph Laplacian.
v) MCFS: is the Multi-Cluster Feature Selection method which has been shown to outperform other methods that preserve the cluster structure of the data.
vi) GreedyFS: The basic greedy algorithm presented in this paper (using recursive update formulas for ***f*** and ***g*** but without random partitioning).
vii) PartGreedyFS: The partition-based greedy algorithm.

Table 1: Summary of previously used algorithms

S. NO.	Name of Paper	Year of Publication	Author	Abstract / Conclusion
1	An Efficient Greedy Method for Unsupervised Feature Selection.	2011	Ali Ghodsi,Mohamd S. Kamel and Ahmed K. Farahat	* This paper proposes a novel method for unsupervised feature selection, which efficiently selects features in a greedy manner. *greedy algorithm is based on an efficient recursive formula for calculating the reconstruction error.
2	An automaticapproachforontology-basedfeatureextractionfrom Heterogeneoustextualresources	2012	Carlos Vicient n, DavidSa´nchez, AntonioMoreno	*pre-annotated inputs in which text has been mapped to their formal semantics according to one or several knowledge structures (e.g. ontologies, taxonomies). * domain-independent, automatic and unsupervised method to detect relevant features from heterogeneous textual resources, associating them to concepts modelled in a background ontology.

3	Greedy Column Subset Selection for Large-scale Data Sets	2013	Mohamed S. Kamel, Ahmed K. Farahat, Ahmed Elgohary, Ali Ghodsi	* data analysts to understand the insights of the data and explore its hidden structure. *This paper presents a fast and accurate greedy algorithm for large-scale column subset selection. *This paper also presents an accurate and ecient MapReduce algorithm for selecting a subset of columns from a massively distributed matrix.
4	A Fast Greedy Algorithm for Generalized Column Subset Selection	2013	Ahmed K. Farahat, Ali Ghodsi, and Mohamed S. Kamel	* the selection of a few columns from a source matrix A that best approximate the span of a target matrix B * We define a generalized variant of the column subset selection problem and present a fast greedy algorithm called Greedy Generalized CSS for solving it.

III. MOTIVATION

There are some reasons that why we need dimension reduction over data instance because:

- It is easy and expedient to accumulate data from sources and Data is not collected only for data mining.
- Data accumulates in an unprecedented speed.

Hence data preprocessing plays significant part for effective machine learning and data mining applications. Dimensionality reduction is an effectual approach to downsizing data.

- The majority machine learning and data mining techniques may not be helpful for high dimensional data and Query correctness and efficiency disgrace speedily as the dimension increases.
- Data compression: efficient storage and retrieval.
- Noise removal: positive effect on query accuracy.

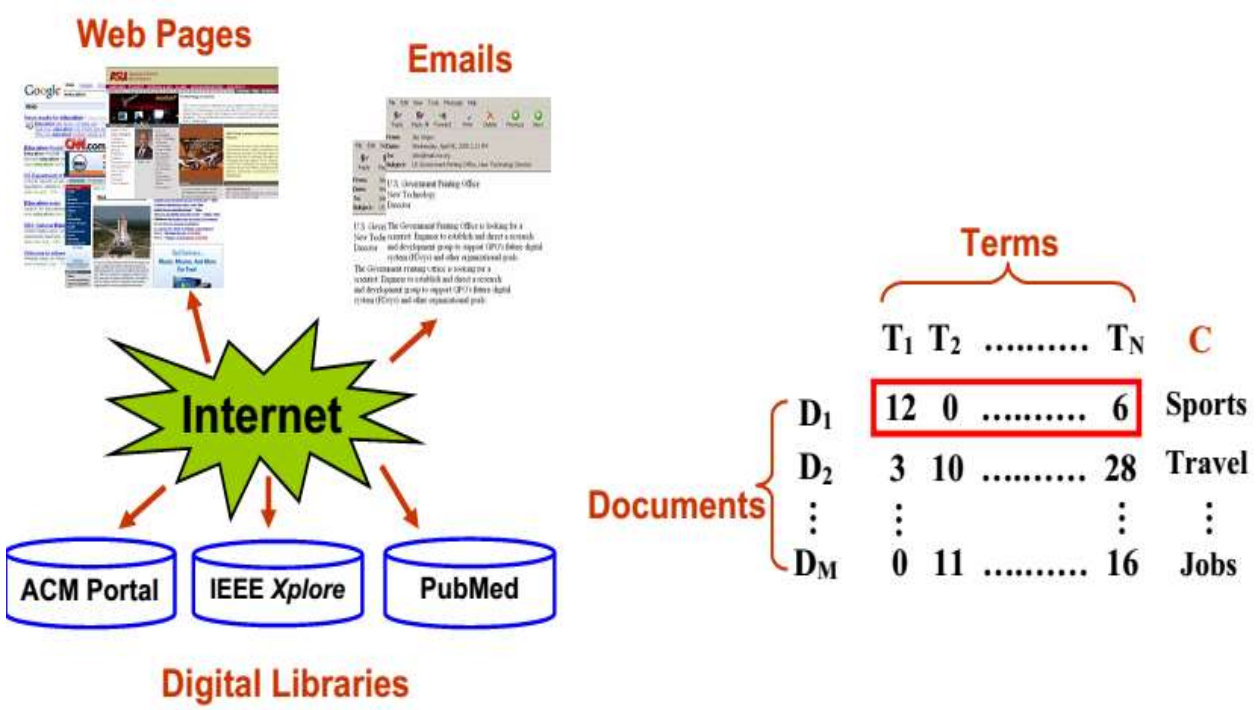

Fig.-1. Document Classification

In above document Classification job is to classify unlabeled documents into categories and the bottleneck is there are thousands of terms, all are not related to job hence we need to apply dimensionality reduction so as to improve the performance of desired data mining application. There are several applications of Dimensionality Reduction:

- Customer relationship management
- Text mining
- Image retrieval

- Microarray data analysis
- Protein classification
- Face recognition
- Handwritten digit recognition

IV. PROBLEM IDENTIFICATION

There are some bottlenecks in dimension reduction approach.

- Physically tagging of huge amounts of training data is very prolonged; furthermore, it is hard for one data mining system to be ported across different domains. Due to the limitation of supervised methods, some semi-supervised approaches have been recommended.
- Order selection and discriminative label identification.
- The intrinsic dimension.
- Data compression for data instance storage.
- Speed of learning
- Predictive accuracy
- Simplicity and comprehensibility of mined result

Problem On Feaure Selection

i. **Low-rank Matrix Approximation:** Given an m _ n matrix A, the rank of A is defined as the maximum number of linearly independent rows or columns. The problem of finding a low-rank approximation of a matrix is defined as: Problem (Low-rank Matrix Approximation) Given an m _ n matrix A and a positive integer k, and an m * n matrix ~A such that:

A~ = arg min[] B; rank(B)_k where k:kF is the Fresenius norm of a matrix.

The objective function of above problem quantifies the discrepancy between the data matrix A and its low-rank approximation B. This discrepancy is referred to as the approximation error or the reconstruction error of the data matrix. In this objective function, the Fresenius norm of the discrepancy matrix is used to quantify the reconstruction error. Other types of norms such as the spectral norm can also be used to quantify this error.

ii. **Data Clustering:** Data clustering is an unsupervised learning task which aims at organizing data instances into groups based on their similarity. Data instances are usually repres nted as points in a multidimensional space of features, and the similarity between these data instances is calculated based on their closeness in that space.

iii. **Hierarchical Clustering:** Hierarchical algorithms for data clustering construct a hierarchy of nested clusters. The top cluster in the hierarchy contains all data points while each cluster at the bottom contains a single data point (also called singleton clusters). The data points of each intermediate cluster are divided into a set of sub-clusters (usually two). The output of a hierarchical clustering algorithm can be graphically represented as a tree.

V. COMPARISON

Table 2: Comparison of algorithms

S.No.	Author/Year	Name of Algorithm	Advantage	Disadvantage
1.	Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu. Clustering-guided sparse structural learning for unsupervised feature selection. IEEE TKDE, 26(9):2138–2150, Sept 2014	CGSSL	Provides label information for the structured learning in optimized form	Feature correlations are not investigated explicitly
2.	Haichang Li ; Inst. of Autom., Beijing, China ; Shiming Xiang ; Zisha Zhong ; Kun Ding Multicluster Spatial–Spectral Unsupervised Feature Selection for Hyperspectral Image Classification IEEE 2015	Unsupervised Spatial-Spectral Feature Selection Method	Best relevant features from hyper spectral image dataset are obtained with approximation	Not applicable for large datasets
3.	Padungweang, P. Padungweang, P. A Discrimination Analysis for Unsupervised Feature Selection via Optic Diffraction Principle IEEE 2012	Unsupervised Feature Selection Via Optic Diffraction Principle	The notion of physical optics is used effectively for discrimination calculation of distribution	Sometimes depends on probability density estimation which requires future search for finding optimal solution

4.	Ahmed K. Farahat Ali Ghodsi Mohamed S. Kamel An Efficient Greedy Method for Unsupervised Feature Selection IEEE 2011	Greedy Method for Unsupervised Feature Selection	Algorithm optimizes a feature selection criterion which measures the reconstruction error of the data matrix based on the subset of selected features	Less efficient for very large data instance.

VI. CONCLUSION

The idea of having explicit semantic information that can be used by intelligent agents in order to solve complex problems of Information Retrieval and Question Answering and to semantically analyze and catalogue the electronic contents. The novel reduction technique will improve performance of SVM data input set to convert data from textual level to conceptual level. In present, accuracy and fast computation is always in demand. In this paper we have discussed diverse feature selection algorithms so as to find out factors which affect the performance of existing algorithm so that we can move further for researching another novel method for data mining application

REFERENCES

[1] Shiming Xiang ; Zisha Zhong ; Kun Ding Multicluster Spatial–Spectral Unsupervised Feature Selection for Hyperspectral Image Classification IEEE 2015.

[2] P.Miruthula1, S.Nithya Roopa Unsupervised Feature Selection Algorithms: A Survey IJSR 2015.

[3] Wee-Hong Ong, Leon Palafox, Takafumi Koseki nvestigation of Feature Extraction for Unsupervised Learning in Human Activity Detection Volume 2, Number 1, pages 30–35, January 2013.

[4] Liang Du, Yi-Dong Shen Unsupervised Feature Selection with Adaptive Structure Learning 2015.

[5] Ahmed K. Farahat Ali Ghodsi Mohamed S. Kamel An Efficient Greedy Method for Unsupervised Feature Selection 2011 11th IEEE International Conference on Data Mining

[6] Yi Yang1, Heng Tao Shen1, Zhigang Ma2, Zi Huang1, Xiaofang Zhou l2,1-Norm Regularized Discriminative Feature Selection for Unsupervised Learning Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence 2015.

[7] L.J.P. van der Maaten * , E.O. Postma, H.J. van den Herik Dimensionality Reduction: A Comparative Review MICC, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. 2015.

[8] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu. Clustering-guided sparse structural learning for unsupervised feature selection. IEEE TKDE, 26(9):2138–2150, Sept 2014.

[9] Jennifer G. Dy Feature Selection for Unsupervised Learning School of Electrical and Computer Engineering US 2003.

[10] Padungweang, P. Padungweang, P. A Discrimination Analysis for Unsupervised Feature Selection via Optic Diffraction Principle IEEE 2012.

[11] Ahmed K. Farahat, Ali Ghodsi, and Mohamed S. Kamel A Fast Greedy Algorithm for Generalized Column Subset Selection 2013.

[12] Jiliang Tang and Huan Liu"Unsupervised feature selection framework for social media data"IEEE trans on knowledge engg and datamining., vol 26, no.12,Dec 2014.

[13] Zechao Li, Jing Liu, Yi Yang, Xiaofang Zhou, Senior Member, IEEE, and Hanqing Lu, Senior Member,IEEE" Clustering-Guided Sparse Structural Learning For Unsupervised Feature Selection"IEEE trans on knowledge engg and data mining., vol 26,sept2014.