www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 8 August 2017, Page No. 22276-22283
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i8.08

High Dimensional Graphical Data Reduction

Smita J.Khelukar[1]

[1]Computer Engineering Department,
SVIT COE, Chincholi,
Sinner, Nasik, Pune University, Maharashtra, India.
smitakhelukar11@gmail.com

Abstract
The coming century is surely the century of data. A combination of blind faith and serious purpose makes our society invest massively in the collection and processing of data of all kinds, on scales unimaginable until recently. In spite of the fact that graph embedding has been an intense instrument for displaying data natural structures, just utilizing all elements for data structures revelation may bring about noise amplification. This is especially serious for high dimensional data with little examples. To meet this test, a novel effective structure to perform highlight determination for graph embedding, in which a classification of graph implanting routines is given a role as a slightest squares relapse issue. In this structure, some preprocessing techniques for instance selection are used. Classification, Clustering are used for accuracy calculation.

Keywords- High dimensional data, Instance selection,Feature Selection,Classification,Clustering.

1. Introduction

The world continues to generate quintillion bytes of data daily, leading to the pressing needs for new efforts in dealing with the grand challenges brought by Big Data. Today, there is a growing consensus among the computational intelligence communities that data volume presents an immediate challenge pertaining to the scalability issue. However, when addressing volume in Big Data analytics, researchers in the data analytics community have largely taken a one-sided study of volume, which is the Big Instance Size factor of the data. The flip side of volume which is the dimensionality To lighten this, one conceivable methodology is to change high To lighten this, one systematically searching through a high-dimensional space, the apparent intractability of accurately approximating a general high-dimensional function, the apparent intractability of integrating a high-dimensional function.

Two of the most influential principles in the coming century will be principles originally discovered and cultivated by mathematicians: the blessings of dimensionality and the curse of dimensionality. The curse of dimensionality is a phrase used by several subfields in the mathematical sciences; I use it here to conceivable methodology is to change high dimensional data into a lower dimensional representation while safeguarding the inborn data structures. This is dimensionality reduction. Inherent data structures can have both nearby and worldwide properties, contigent upon the applications. Nearby properties frequently allude to the nearby neighborhood relationship for example in LPP, while illustrations of worldwide properties incorporate class detachment in LDA, the worldwide change in PCA, and the worldwide most brief way between any sets of data tests in the Isomap technique.

refer to the apparent intractability of systematically searching through a high-dimensional space, the apparent intractability of accurately approximating a general high-dimensional function, the apparent intractability of integrating a high-dimensional function.

The blessings of dimensionality are less widely noted, but they include the concentration of measure phenomenon, which means that certain random fluctuations are very well controlled in high dimensions and the success of asymptotic methods,

used widely in mathematical statistics and statistical physics, which suggest that statements about very high-dimensional settings may be made where moderate dimensions would be too complicated.

2. Literature Survey

Numerous issues in data preparing include some type of dimensionality lessening. Locality Preserving Projection (LPP) [3] is direct projective maps that emerge by unraveling a variational issue that ideally protects the area structure of the dataset. LPP ought to be seen as a distinct option for Principal Component Analysis (PCA) - an established straight method that activities the information along the bearings of maximal fluctuation. At the point when the high dimensional information lies on a low dimensional complex installed in the encompassing space, the Locality Finding so as to preserve Projections are acquired the ideal direct approximations to the Eigen functions of the Laplace Beltrami administrator on the complex. Thus LPP offers a large portion of the information representation properties of nonlinear strategies for example Locally Linear Embedding. Yet LPP is straight and then some critically is characterized all over the place in encompassing space as opposed to simply on the preparing information focuses.

Volumes of high dimensional information [4] for example worldwide atmosphere designs, stellar spectra or human quality conveyances, frequently face the issue of dimensionality diminishment: pending important low dimensional structures covered up in their high dimensional perceptions. Here portray a way to deal with tackling dimensionality diminishment issues that uses effortlessly measured nearby metric data to take in the hidden worldwide geometry of an information set. Not at all like established systems, for example central part investigation (PCA) and multidimensional scaling (MDS)[5] ,the methodology is fit for finding the nonlinear degrees of flexibility that underlie complex common perceptions for example a face under distinctive review conditions. As opposed to past calculations for nonlinear dimensionality diminishment, own efficiently processes an all inclusive ideal arrangement, what's more for an imperative class of information manifolds is ensured to unite asymptotically to the genuine strum.

Locally Straight Implanting (LLE) [7] an unsupervised learning calculation that processes low dimensional, neighborhood protecting embedding of high dimensional inputs. Not at all like grouping techniques for neighborhood dimensionality lessening, LLE maps its inputs into a solitary worldwide direction arrangement of lower dimensionality and its advancements don't include nearby minima. By abusing the neighborhood symmetries of straight reconstructions, LLE's ready to take in the worldwide structure of nonlinear manifolds for example created by pictures of confronts or records of content.

3. Proposed System

The immense growth of feature dimensionality in data analytics has exposed the inadequacies of many computational intelligence methodologies that exist to date. Hence there is an urgent need for the conception of new paradigms and methodologies that can cope with the emerging phenomenon of Big Dimensionality. Correspondingly, how to solicit the key features to concisely represent the data and the prediction model well, while facilitating fast prediction and reduced storage, are among the important tasks of Big Data analytics.

We will consider what statisticians consider the usual data matrix, a rectangular array with *N* rows and *p* columns, the rows giving different *observations* or *individuals* and the columns giving different *attributes* or *variables*. There are broad range of applications where we can have *N* by *p* data matrices.
For example:

- **Web –Term Document Data:**
  In this model, one compiles *term-document matrices*, *N* by *p* arrays, where *N*, the number of documents, is in the millions, while *p*, the number of terms (words), is in the tens of thousands, and each entry in the array measures the frequency of occurrence of given terms in the given document, in a suitable normalization.
- **Sensor Array Data:**
  An array of *p* sensors is attached to the scalp, with each sensor records *N* observations over a period of seconds, at a rate of X thousand samples, second.
- **Gene Expression Data:**
  Data on the relative abundance of *p* genes in each of *N* different cell lines.
- **Imagery:**
  We can view a database of images as an *N*-by-*p* data matrix. Each image gives rise to an observation; if the image is *n* by *n*,

then we have $p = n2$ variables. Different images are then our different individuals.

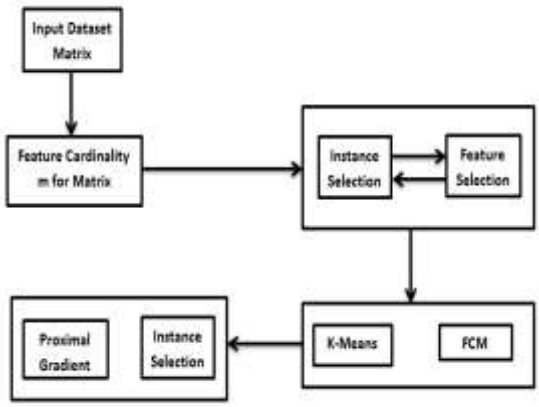


Figure 1: System Architecture

As shown in Figure1, take input dataset matrix of high dimensional data. The architecture diagram shown gives the clear view of the system.

3.1 Dimensionality Reduction Techniques:

3.1.1 Instance Selection:

Instance selection is an important data pre-processing step that can be applied for reducing original dataset to manageable volume. In it, zero reduction, variance reduction and similarity reduction is performed. Zero reduction involves eliminating column from dataset having number of zero (%) more. In variance reduction, variance of columns calculated. Having variance of column more than 0.50 are eliminated. Similarity reduction involves elimination of columns having more similarity among them.

3.1.2 Feature Selection:

As data contains many features that are either redundant or irrelevant. So removing that features does not incur loss of information.

3.1.3 Classification:

In classification, one of the p variables is an indicator of class membership. Many approaches have been suggested for classification, ranging from identifying hyperplanes which partition the sample space into non-overlapping groups, to k-nearest neighbor classification. Train classifier to classify features. Select most active features. Calculate accuracy with and without PCA. Classification.

3.2 Data Selection:

Cluster Analysis could be considered a field all its own, part art form, part scientific undertaking. One seeks to arrange an unordered collection of objects in a fashion so that nearby objects are similar. There are many ways to do this, serving many distinct purposes, and so no unique best way. An obvious application area would be in latent semantic indexing, where we might seek an arrangement of documents so that nearby documents are similar and an arrangement of terms so that nearby terms are similar.

3.2.1 K-Means:

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. In particular, the parameter k is known to be hard to choose (as discussed above) when not given by external constraints. Another limitation of the algorithm is that it cannot be used with arbitrary distance functions or on non-numerical data.

3.2.2 FCM:

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. As already told, data are bound to each cluster by means of a Membership Function, which represents the fuzzy behavior of this algorithm. To do that, we simply have to build an appropriate matrix named U whose factors are numbers between 0 and 1, and represent the degree of membership between data and centers of clusters.

3.3 Merge two dataset:

If any user want data from two datasets then it becomes difficult to process both the dataset for reduction. So applying instance selection on both datasets reduce size of datasets. Perform merge operation on both dataset to form only one dataset. Dimensionality of merge dataset is much lower than input datasets.

DOI: 10.18535/ijecs/v6i8.08

4. Algorithm

4.1 Sparse Graph Embedding Algorithm for Feature Selection:

The optimization problem has a combinatorial number of constraints. However, only a few of them are active. Exploiting this observation, we adopt the cutting plane algorithm to solve the QCQP problem. The cutting plane algorithm iteratively finds the most active constraint.

Input: data X ϵ $R^{d\times n}$ a positive semi-definite matrix S, the desired feature cardinality m.
(1) Initialize π = Ø and compute T according to (3). Assign t := 1.
(2) Iterate the following two steps until convergence.
(a) Update V by solving the sub problem.
(b) Find the most active constraint, which is indicated by p^t, by solving p^t= argmaxp f(V,p); based on V .
Update π by π: = π ∪ $\{p^t\}$ and t by t := t+1;
Output: π = $\{p^1 p^2 \ldots . p^k\}$, with each p^i indexing the selected features

4.2 The Most Active Constraint Selection:

The most active constraint can be identified by choosing the features with the m highest values in s. The most active constraint obtained is then added to the active constraint.
Input: Data X ϵ $R^{d\times n}$, dual variable V , the desired number of Features m, and the selection vector p.
(1) Set all the entries of p to 0.
(2) Compute si = $\sum_{j=1}^{k}$ $(Ai,j)^2$, ∀ = 1,……d.
(3) Sort s in descending order.
(4) Set m entries of p w.r.t. the top m values of s.
Output: p which defines the most active constraint.

4.3 Moreau Projection Algorithms:

After updating the active constraint set P, we then solve the subproblem with reduced constraints as defined by P. Since the number of constraints in P is no longer large, this problem is readily solved by a sub-gradient method, such as simple MKL.
However, solving this problem w.r.t. the dual variables V can be very expensive, in particular when n is very large. Assume there are k active constraints in P. Even though there are a large number of features in X, at most mk features are chosen by P. Based on this observation, the subproblem might be solved more efficiently w.r.t. the primal variables W.

Moreau Projection: Sι()G
Input G= [G1,G2,...,Gk] and s=1/ι.
(1) Calculate ˆ ut = ||Gt||F for all t=1,...,k.
(2) Sortˆ u to obtain u such that u(1) ≥ ... ≥ u(k).
(3) Find ρ = max{t|ut −s 1 + tsPt i=1 ui > 0,t = 1,...k}.
(4) Calculate the threshold value S = s/1 + ρsPρ i=1 ui.
(5) Compute o = soft (ˆu, S).
(6) Compute and output : Sι(G).

4.4 Accelerated Proximal Gradient Algorithm:

Given an ultrahigh dimensional sparse data matrix, removing the data mean (zero-centering) could make the matrix very dense. The data matrix can be used instead for regression to remove the data offset. As for the proposed framework, zero-centering can be performed in each subproblem.

Initialization: Initialize the lipschitz constant Lt = Lt−1 and set Ω−1 = Ω0 by warm start, ι0 = Lt, n ∈ (0, 1), parameter `−1 = `0 = 1, and k=0.

(1) Set V k = Ωk + (ek−1 −1)/ek(Ωk −Ωk−1).
(2) Set ι = nιk. Repeat Set G=V k −1/ιOf(V k), compute Sι(G). if F(Sι(G)) ≤ Q(Sι(G),V k), set ιk = ι, stop ,break; else ι= min {n−1ι,Lt}. End Until convergence F(Sι(G)) ≤ (Sι(G),V k)
(3) Set Ωk+1 = (Sιk(G)).
(4) Let %k+1 = (1 +p (1 + 4(%k2))/2. Let k=k+1
(5) Quite if the stopping condition is achieved. Otherwise go to, step 1.
(6) Let Lt = n2ιk and return.

4.5 Principal Component Analysis (PCA):

Principal Component Analysis depends on big set of data to analyze that set in terms of relationship between the individual points in that data set. As dealing with high dimensional data, calculate covariance matrix. Eigenvectors and eigenvalues are finding for covariance matrix. So that is way of identifying patterns in a dataset.

(1) Mean Center the data.
(2) Compute the covariance matrix of the dimensions.

(3) Find the eigenvectors of covariance matrix.
(4) Sort the eigenvectors in decreasing order of eigenvalues.
(5) Project onto eigenvectors in order.

4.6 K-Means Clustering:

K-means algorithm is one of the simplest unsupervised learning algorithms that partition technic vectors into k clusters so that the within group sum of squares is minimized. K-means clustering technic is a method of vector quantization originally from signal processing that is popular for cluster analysis in data.mining. K-Means follows a simple way to classify a given dataset.

Let $X = \{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots,v_c\}$ be the set of centers.

(1) Randomly select '*c'* cluster centers.

(2) Calculate the distance between each data point and cluster centers.

(3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

(4) Recalculate the new cluster center.

(5) Recalculate the distance between each data point and new obtained cluster centers.

(6) If no data point was reassigned then stop, otherwise repeat from step (3).

4.7 FCM Clustering:

In FCM, fuzzy membership is calculated. Each data point is assigned to fuzzy membership corresponding to each cluster on the basis of distance between the cluster center and data point.
Let $X = \{x_1, x_2, x_3 ..., x_n\}$ be the set of data points and $V = \{v_1, v_2, v_3 ..., v_c\}$ be the set of centers..

(1) Randomly select *'c'* cluster centers.

(2) Calculate the fuzzy membership.

(3) Compute the fuzzy centers.

(4) Repeat step (2) and (3) until the minimum J value is achieved

5. Mathematical Model

A mathematical model is a description of a system using mathematical concepts and language. Mathematical model used to maximize a certain output. The system under consideration will require certain inputs. The system relating inputs to outputs depends on other variables defined in the below section with the help of Venn Diagram as shown in Figure.2.

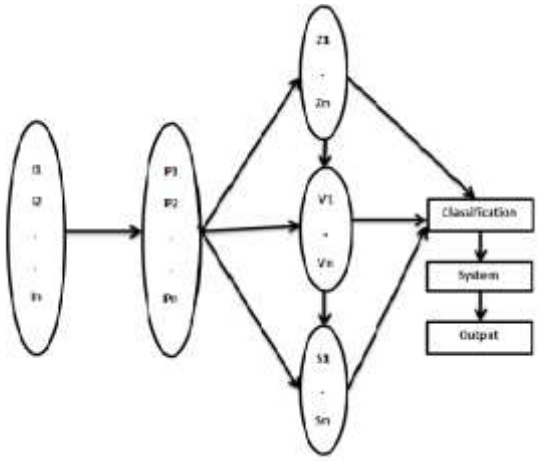


Figure 2: Functional Dependency Of System

Set Theory:
Let I is a set of input i.e. Dataset, R be the set of Reduction Techniques
F is the set of functions used for the implementation.
O is the output.
S= (I, F, O)
I: Input Dataset
R: Set of Reduction Techniques F: Set of Functions
O: Set of Output
F=F1,F2,.....F10.

F1: Loading dataset.
F2: Finding zero columns in dataset.
F3: Finding variance of column in dataset.
F4: Checking similarity among columns of dataset.
F5: PCA Classification.
F6: Merge dataset.
F7: Applying K-Means Clustering.
F8: Calculate Accuracy of dataset with or without reduction after K-Means.
F9: Applying FCM Clustering.

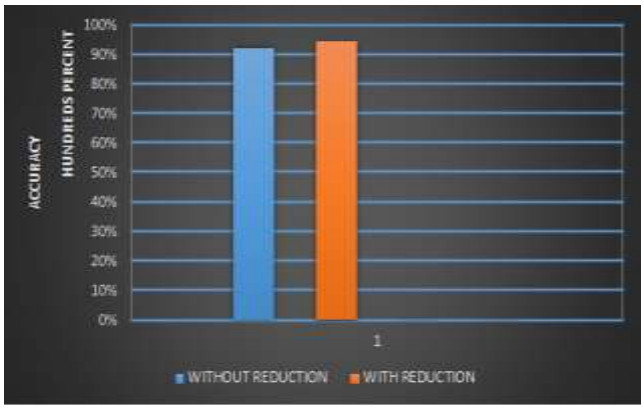
F10: Calculate Accuracy of dataset with or without reduction after FCM.

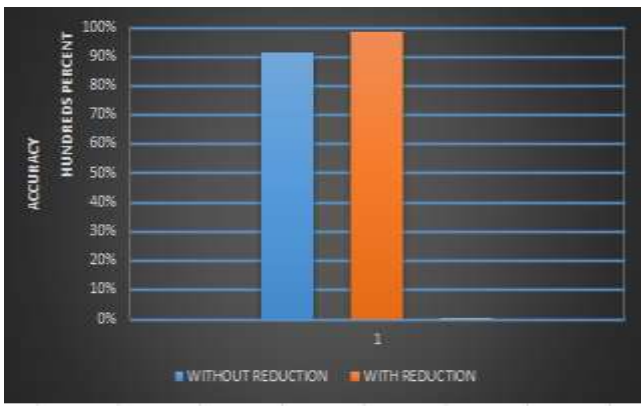
Let I be set of input dataset and O be the set of output.
Let R be the set of different reduction techniques which can be applied on input dataset to get corresponding output.
O = {I, Z, V, S, R}
Where I = {I1, I2, …,In, In≠0}, I is set of input string
Where Z = {Z1, Z2, …,Zn, Zn≠0}, Z is set of zero column.
Where V = {V1, V2,…..,Vn,Vn≠0}, V is set of variance of column.
Where S = {S1, S2 ,…,Sn,Sn≠0},S is the set of similarity of column.
Where R = {R1,R2,,…..,Rn,Rn≠0}, R is set of Reduction Techniques.
Let Fz (I) → Z Where Fz is a function that takes the input dataset and provides the zero column of the dataset.
Let Fv (P) → V Where Fv is a function that takes the input dataset having zero reduction and provides variance reduction of the dataset.
Let Fs (P) → S Where Fs is a function that takes the variance reduction dataset and provides similarity reduction dataset.
Let Fr (P) →R Where Fr is a function that takes the dataset and provides reduction form of result.
X: Importing the dataset.
F(X): As per the users requirement the dataset is uploaded in the system in proper format .
F1: Database updation as per the user need.
X: Applying reduction techniques.
F(X): What can be the probable techniques for reduction dimensionality.
F3: Techniques are find out and weights are assigned to the techniques.

X: Input dataset.
F(X): The input dataset is parsed on the basis of parameters like rows, attributes etc.
F2: Reduce the input by finding zero columns.
X: Dataset is given as input for zero reduction.
F(X): Dataset is given as an input for finding variance.
F3: Reduce the input by finding variance of column.
X: Searching and finding the similar columns in dataset is done.
F(X): If entered dataset having similarity among columns of dataset then similarity column reduction is performed.
F4: Obtaining results of classification.
X: Classification on reduction dataset is performed.
F(X): Classification is performed depending on reduction factor.
F7: Applying Merge for input dataset.
X: K-Means Clustering.
F(X): After classification, K-Means algorithm is applied for getting the accurate results.
X: FCM Clustering.
F(X): After classification, FCM algorithm is applied for getting the accurate results.
F9: Analysis is done by graph i.e accuracy estimation is calculated here.

6. Result Analysis and Discussion

In existing work, Feature selection and Graph embedding tasks have been done independently or mutually exclusively. This paper instead proposes a novel paradigm to unify these two schemes by performing Feature selection and Graph embedding simultaneously. Classification, Clustering, etc. preprocessing dimensionality reduction techniques are applied on high dimensional datasets so time complexity of System get reduced to much extent. Accuracy and Efficiency of required result get improve.





Both Graph shows result for wine dataset. Using K-Means and FCM clustering. As we use reduction techniques so no.of features get reduced to much extent. So it helps to improve accuracy of classification using Naïve Baye's and ID3 to get higher accurate and efficient output data in low dimensional form from high dimensional data.

7. Conclusion

In this system user is provided with extra facilities of dataset dimensionality reduction. User is allowed to search specific data from large amount of data. Mainly if the user is from non-technical background and has very few details of dataset then its easy to use this system. Another feature of this system is that reduction results are very accurate. Also the main concentration is not only on accuracy but also on efficiency .This is an important feature of this system even if the datasets increases the efficiency does not degrade. The main aim was to achieve efficiency along with the accuracy. The experimental result shows that even if the dataset increases the reduction time does not increase much. The reduction results are found in effective time only. This system is especially helpful for the banking, education, industrial, business purposes where there are more chances of high dimensional data. User can also merge two datasets after reducing dimensionality of two datasets.

References

1. Marcus Chen, Ivor W. Tsang, Mingkui Tan, and Tat Jen Cham,"A Unified Feature Selection Framework for Graph Embedding on High Dimensional Data", IEEE Trans.on Knowledge and Data Engg VOL. NO.6, JUNE 2015.
2. Y.Zhai, Y. Ong and I. Tsang,"The emerging "Big Dimensionality"" , IEEE Comput.Intell.Mag. VOL. NO.9, NO.3, pp.14-26, JULY 2014.
3. X.He and P.Niyogi,"Locality Preserving Projections", in Proc.Adv. Neural Inf.Process. Syst., VOL NO.16, 2004, p.153.
4. S.T.Roweis and L.K.Saul,"Nonlinear Dimensionality Reduction by Locally Linear Embedding", Science, Vol.290, no.5500,pp.2323-2326, 2000.
5. I.M. Johnstone and A.Y. Lu,"On consistency and sparsity for Principal Component Analysis in High Dimensions", J. Am.Statist.Assoc., Vol. 104, no.486, pp.682-693, 2009.
6. Q.Gu, Z.Li, and J.Han,"Generalized Fisher Score for Feature Selection" In Proc.27th Conf. Uncertainty Artif. Intell. 2011, pp.266-273.
7. V.Q.Vu, J.Cho, J.Lei and K. Rohe,"Fantope Projection and Selection: A near-Optimal Convex Relaxations of Sparse PCA", in Proc. Adv. Neural Inf.Process Syst., 2013, pp.2670-2678.
8. C.Hou, F.Nie, X.Li, D.Yi and Y.Wu,"Joint Embedding Learning and Sparse Regression: A Framework for Unsupervised Feature Selection", IEEE Trans. Cybern., Vol.44, NO.6, pp.793-804, JUNE 2013.
9. X. Cai, F. Nie, and H. Huang, Exact top-k feature selection via l2,0- Norm Constraint, in Proc.23rd Joint Conf. Artif. Intell.,2013, pp.1240- 1246.
10. F.Nie, H.Huang, X. Cai, and C.Ding,"Efficient and Robust Feature Selection via Joint 12,1- Norms Minimization" Adv. Neural Inf. Pro. Syst., vol.23, pp-1813-1821, 2010.
11. S.Xiang, F.Nie, G. Meng, C.Pan, and C.Zhang,"Discriminative Least Squares Regression For Multiclass Classification and Feature Selection", IEEE Trans.Neural Netw.Learn Syst,vol.23,no.11,pp.1738-1754,Oct.2012.
12. D.Cai, C.Zhang and X.He,"Unsupervised Feature Selection For Multi- Cluster Data", in Proc. 16th ACM SIGKDD Int.Conf. Knowledge Discovery Data Min., 2010, pp.333-342.
13. F.Bach, S.D. Ahipasaoglu and A.dAspremont,"Convex Relaxations For Subset Selection", arXiv preprint arXiv:1006.3601,2010.
14. Y.Liu, F.Nie,J. Wu and L. Chen,Efficient Semi Supervised Feature Selection with

Noise Insensitive Trace Ratio Criterion, Neurocomput- Ing, vol.105, pp.12-18, 2013.

15. M.Tan, L.Wang and I.W. Tsang,"Learning Sparse SVM for Feature Selection on Very High Dimensional Datasets, in Proc.27th Int.Conf. Mach.Learn, 2010, pp.1047-1054
16. M. Tan, I.W. Tsang and L.Wang,"Towards Ultrahigh Dimensional Feature Selection for Big Data", J. Mach. Learning Research, Vol.15, pp.1371-1429, 2014.
17. R.Bellman and R. Bellman (1957). Dynamic Programming. Ser.P (Rand Corporation). Princeton, NJ, USA: Princeton University Press. [Online]Available: http://books.google.com.sg/books?id=rZW4ugAACAAJ

Author Profile

Ms. Smita J.Khelukar has completed her B.E.in Computer Engineering from Pune University and currently pursuing Masters of Engineering from SVIT, Chincholi, Nashik, India.