Reduction Of High Dimensional Graphical Data

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

Mukund.B.Wagh[2]
[2]Computer Engineering Department
SVIT COE, Chincholi
Sinner, Nasik, Pune University, Maharashtra, India.
mukund.wagh.81@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, a twofold component selector is acquainted with normally handle the component cardinality at all squares detailing. The proposed strategy is quick and memory proficient. The proposed system is connected to a few graph embedding learning issues, counting administered, unsupervised and semi supervised graph embedding.

Keywords- Feature selection; High dimensional data; Sparse graph embedding; Sparse principal component analysis;Subproblem Optimization;

I. INTRODUCTION

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 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.

To lighten this, one 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.

Numerous feature selection strategies have been proposed in diverse learning settings with diverse component significance measures. These strategies can be arranged into two classes, to be specific, the regulated and unsupervised routines. For the regulated routines there are two principle highlight significance measures, distance based measures and the connection based measures. In particular, the separation based measures characterize the critical components as those that different classes better and group the inside of class tests for example LDA based component determination routines. In relationship based component choice methods the critical components are those that relate well with class names furthermore give better forecast results. In the

unsupervised techniques because of the nonattendance of class marks a few criteria have been proposed to assess the component significance taking into account diverse learning settings for example information measure, fluctuation measure and region measure.

II. 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.

III. 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.

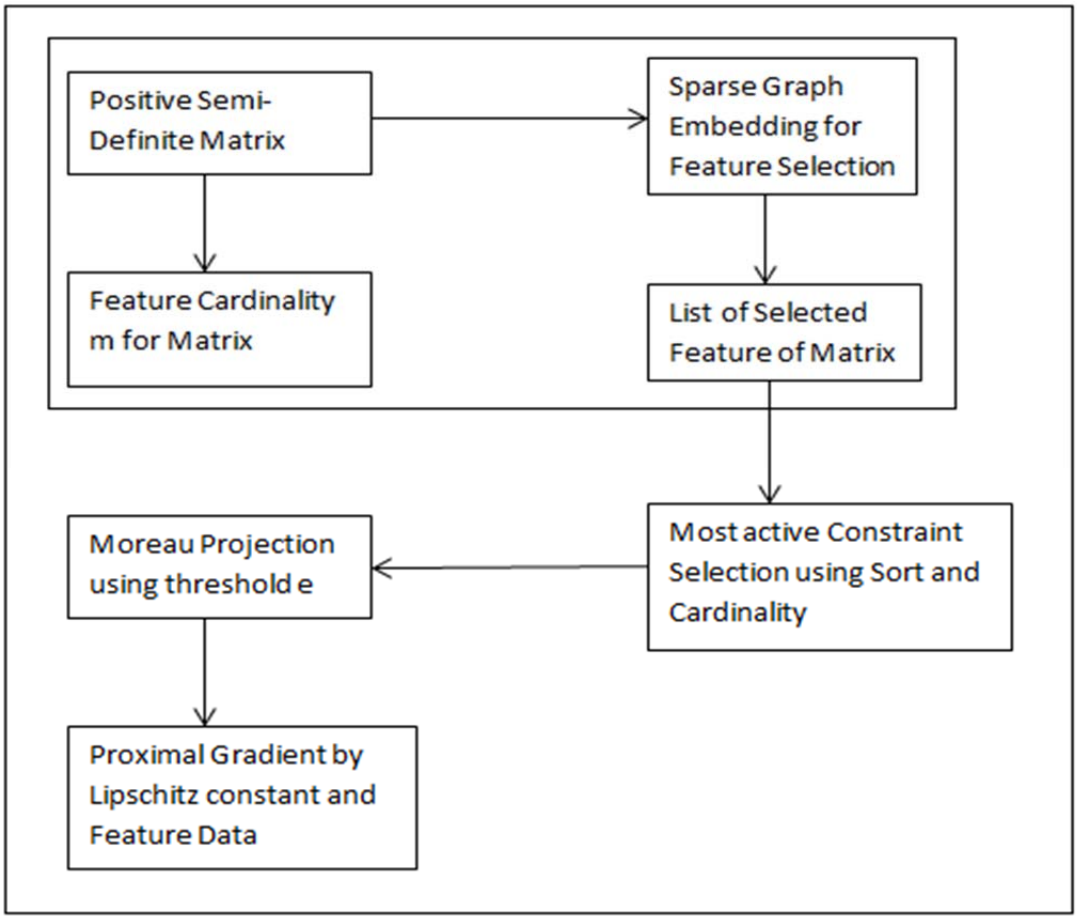

Figure 1. System Architecture

As shown in Figure1, take positive semi-definite matrix of high dimensional data. The architecture diagram shown gives the clear view of the system. By misusing the minimum squares of detailing chart inserting, introduce a binary feature selector with specifically oblique the coveted number of features. Then further reformulate the resultant issue as a curved semi-infinite programming issue (SIP). This novel feature selection plan can be connected to unsupervised, directed and semi administered learning tasks in saving the relating natural information structures by means of low dimensional embeddings. By exploiting the perception that just a couple of limitations are dynamic in the resultant SIP issue.
So proposed cutting plane technique which basically directs a grouping of accelerated proximal slopes on an arrangement of components just. Distinguish the ideal discriminative and uncorrelated element subset to the yield names means here as support features which realizes about significant upgrades in expectation execution.
During learning process, basic gathering structures of related components connected with every support feature indicated as Affiliated features can likewise be found with no extra cost. These partnered elements serve to enhance the interpretations on the learning tasks.

A. *Sparse Graph Embedding for Feature Selection*:
The optimization issue has a combinatorial number of limitations. In any case, just a couple of them are dynamic. Abusing this perception, the slicing plane calculation to take care of the QCQP issue. The cutting plane calculation iteratively finds the most dynamic constraint, and adds it to the dynamic constraint set, which is introduced to an empty set.

- *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 SparseGE-PCA.

- *Clustering*:

  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.

B. The Subproblem Optimization:

Subsequent to the redesigning the dynamic constraint set , fathom the subproblem with lessen requirements as characterized by. Since the quantity of imperatives in is no more extensive this issue is prominently illuminated by a subgradient technique such as simple MKL. On the other hand, tackling this issue w.r.t. the double variables V can be extremely costly.

C. Handling High Dimensional Sub problems:

Given an ultrahigh dimensional inadequate data matrix, evacuating the data mean (zero focusing) could make the lattice extremely thick. The data matrix can be utilized rather for relapse to evacuate the data balance. With respect to the proposed system, zero focusing can be performed in each subproblem. Zero focusing could likewise influence the calculation of some relapse reactions.

IV. ALGORITHM

A. 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 $\pi = \emptyset$ 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: = $\pi \cup \{p^t\}$ and t by t := t+1;

Output: $\pi = \{p^1 p^2 \ldots . p^k\}$, with each p^i indexing the selected features

B. 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.

C. Moreau Projection Algorithm:

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).

D. 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.

V. 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.

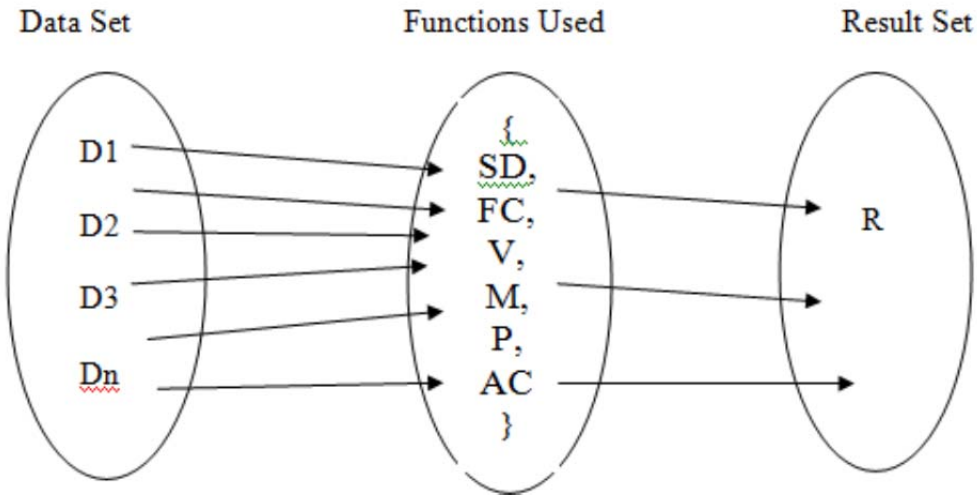

Figure 2. Functional Dependency Of System

Let S is the Whole System Consist of
S= {I, P, O}
Where,
I – input,
P-Procedure,
O- Output

P= {SD, FC, V, M, P, AC}
SD- Positive Semi-Definite Matrix
FC= Feature Cardinality
V= Dual Variable
M= No. of Features
P=Selection Vector
AC= Active Constraints

Set Theory:
1. Identify failure cases as FL:
   Failure occurs when -
   FL = F1, F2, F3
   F1 = f — 'f' error or failure occurs if wrong dataset chosen for keyword search.
2. Identify success case SS:
   Success is defined as-
   SS = S1, S2, S3, S4
   S1 = s — 's' if keyword is in the dataset.

Above theory explains the association between existing search method(raw transactions) and the proposed work, while the following diagram shows that how dataset, functions and result relate with each other.

VI. RESULT ANALYSIS AND DISCUSSION

TABLE I

Dataset Name	Training Sample	Test Sample	No of Features	Dimension Reduction Techniques	Naïve Bayes Accuracy	Feature Reduced
News20	15,935	3,993	62,060	No	63	0
News20	15,935	3,993	58,910	PCA	63	29,455
RCV1	20,242	67,739	47,236	No	58.97	0
RCV1	20,242	67,739	47,236	PCA	54.14	23,618

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.
Table 1 shows result for dataset News20 and RCV1. As we use PCA techniques so no.of features get reduced to much extent. So it helps to improve accuracy of classification task to get higher efficient output data in low dimensional form from high dimensional data.

VII. CONCLUSION

This paper proposes novel unified together system to choose highlights for summed up diagram installing. It uses a component selector to specifically improve highlight subsets for chart inserting in demonstrating the intrinsic data structures, empowering a heartier installing, particularly for high dimensional information with a little specimen size. Its proficiency what's more, viability have been exhibited with a progression of test for grouping, characterization, and perception. In the trials, the proposed strategies beat the present state-of-art calculations for unsupervised, managed, and semi-supervise learning undertaking. The proposed structure showed its computational and memory effectiveness in taking care of ultrahigh dimensional information for order.

ACKNOWLEDGEMENT

It is a great pleasure to acknowledge those who extended their support, and contributed time and psychic energy for the completion of this project work. At the outset, I would like to thank my project guide Prof. M.B.Wagh, who served as sounding board for both contents and programming work. His valuable and skillful guidance, assessment and suggestions from time to time improved the quality of work in all respects. I would like to take this opportunity to express my deep sense of gratitude towards his, for his invaluable contribution in this project work. I am also thankful to Prof. S.M.Rokade, Head of Computer Engineering Department for his timely guidance, inspiration and administrative support without which my work would not have been in process. I am also thankful to the all staff members of Computer Engineering Department and Librarian, SVIT Chincholi, Nasik. Also I would like to thank my colleagues and friends who helped me directly and indirectly in this Project work. Lastly my special thanks to my family members for their support and co-operation during this Project work.

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