ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

An Enhanced Method for Randomized Dimensionality Reduction Using Roughset Based K-Means Clustering

S. Brindha[1] , Dr. Antony Selvadoss Thanamani[2]

M.Phil Research Scholar, Dept of Computer Science, NGM College, Pollachi, Tamil Nadu, India[1]

Assistant Professor and Head, Dept of Computer Science, NGM College, Pollachi, Tamil Nadu, India[2]

ABSTRACT: Dimensionality Reduction is the application of data mining techniques to discover patterns from the datasets. Finding the best features that are similar to a test data is challenging task in current trend. To discover the significance features have more frequent change in the structural information, which involves feature dimensionality reduction, linked to one another and elimination of non-structural information. The proposed research work presents a new approach to measure the features (attributes) in unsupervised datasets using the methodologies namely, preprocessing, k-means based principal component analysis algorithm, Roughset Based Feature Selection and Rough-Set Based K- Means Feature Selection. Data feature selection and dimensionality reduction is characterized by a regularity analysis where the feature values correspond to the number times that term appears in the dataset. This research proposes an enhanced roughset based k-means (*RK*) method to estimate the feature searching is measured using genetic optimization method corresponding unsupervised data. Each feature contains objective function and their description which is used to identify the type of datasets. Initially, the total numbers of features are identified to enhanced *RK* feature selection of the datasets where the terms of match between the features are identified with help of genetic algorithm.

KEYWORDS: Rough-Set, k-means, feature selection, Dimensionality Reduction.

I. Introduction

Clustering is ubiquitous in science and engineering with numerous application domains ranging from bioinformatics and medicine to the social sciences and the web [1]. Perhaps the most well-known clustering algorithm is the so-called “k-means” algorithm or Lloyd’s method [2]. Lloyd’s method is an iterative expectation-maximization type approach that attempts to address the following objective: given a set of Euclidean points and a positive integer k corresponding to the number of clusters, split the points into k clusters so that the total sum of the squared Euclidean distances of each point to its nearest cluster center is minimized. Due to this intuitive objective as well as its effectiveness [3], the Lloyd’s method for k-means clustering has become enormously popular in applications [4]. In recent years, the high dimensionality of modern massive datasets has provided a considerable challenge to the design of efficient algorithmic solutions for k-means clustering. First, ultra-high dimensional data force existing algorithms for k-means clustering to be computationally inefficient, and second, the existence of many irrelevant features may not allow the identification of the relevant underlying structure in the data [5].

The main aim of feature selection (FS) is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. In real world problems FS is a must due to the abundance of noisy, irrelevant or misleading features. For instance, by removing these factors, learning from data techniques can benefit greatly. Given a feature set size *n*, the task of FS can be seen as a search for an \optimal" feature subset through the competing 2*n* candidate subsets. The definition of what an optimal subset is may vary depending on the

DOI: 10.15680/IJIRCCE.2015. 0311067

problem to be solved. Although an exhaustive method may be used for this purpose, this is quite impractical for most datasets.

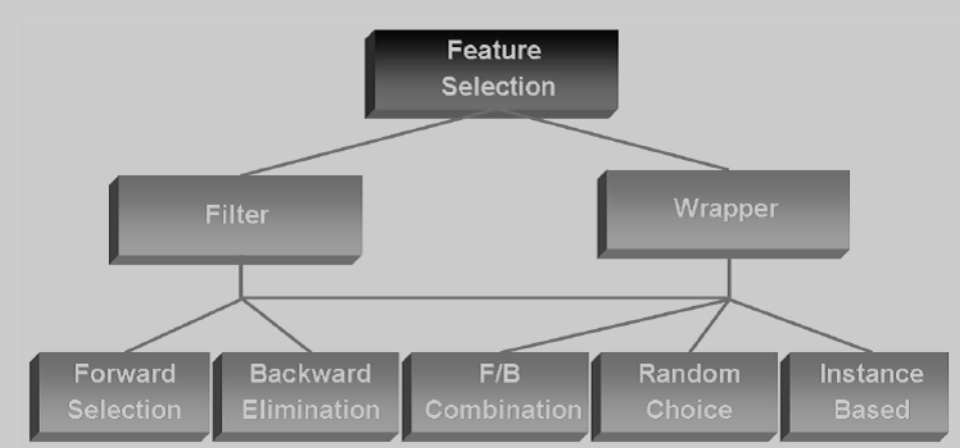

Fig 1. Aspects of feature selection

Cluster analysis is the general task to be solved which means that, it is not one specific algorithm. It is an result of various algorithms itself, in order to be efficient at clustering. It is distinguished by various type of clustering: Hierarchical (nested) versus partitioned (unnested), Exclusive versus overlapping versus fuzzy, complete versus partial. The simple definition of *k-means* clustering, as mentioned earlier, is to classify data to groups of objects based on attributes/features into K number of groups. K is positive integer number. K-means is Prototype-based (center-based) clustering technique which is one of the algorithms that solve the well-known clustering problem. It creates a one-level partitioning of the data objects. K-Means (KM) defines a prototype in terms of a centroid, which is the mean of a group of points and is applied to dimensional continuous space. Another technique as prominent as K-means is K-medoid, which defines a prototype that is the most representative point for a group and can be applied to a wide range of data since it needs a proximity measure for a pair of objects.

II. Related work

In author [2] proposed that in pulse-code modulation (PCM), with a given ensemble of signals to handle, the quantum values should be spaced more closely in the voltage regions where the signal amplitude is more likely asymptotic fractional density of quanta per unit voltage should vary as the one-third power of the probability density per unit voltage of signal amplitudes. In this paper the corresponding result for any finite number of quanta is derived; that is, necessary conditions are found that the quanta and associated quantization intervals of an optimum finite quantization scheme must satisfy. In [3] authors investigated variants of Lloyd's heuristic for clustering high-dimensional data in an attempt to explain its popularity (a half century after its introduction) among practitioners, and in order to suggest improvements in its application. We propose and justify a clusterability criterion for data sets. We present variants of Lloyd's heuristic that quickly lead to provably near-optimal clustering solutions when applied to well-clusterable instances. In [4] proposed the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) in December 2006: C4.5, k-means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART. In [5] illustrated the NIPS 2003 workshops included a feature selection competition organized by the authors. We provided participants with fivedatasets from different application domains and called for classification results using a minimal number of features. The competition took place over a period of 13 weeks and attracted 78 research groups. In [6] discussed the problem of clustering data points. Given n points in a larger set (for example, R/sup d/) endowed with a distance function (for example, L/sup 2/ distance), we would like to partition the data set into k disjoint clusters, each with a "cluster center", so as to minimize the sum over all data points of the distance between the point and the center of the cluster containing the point. The problem is provably NP-hard in some high dimensional geometric settings, even for k=2. In [8] proposed the existence of small coresets for the problems of computing k-median and k-means clustering for points in low dimension. In other words, we show that given a point set P in Rd, one can compute a weighted set S ⊆ P, of size O(k ε-d log n), such that one can compute the k-median/means clustering on S instead of on P, and get an (1+ε)-approximation. [10] proposed an efficient implementation for a k-means clustering algorithm. The novel feature of our algorithm is that it uses coresets to speed up the algorithm. A coreset is a small weighted set of points that approximates the original point set with respect to the considered problem. In [11] presented the k-means method is a widely used clustering technique that seeks to minimize the average squared distance between points in the same cluster. Although it offers no accuracy guarantees, its simplicity and speed are very appealing in practice. By augmenting k-means with a simple, randomized seeding technique, we obtain an algorithm that is O(logk)-competitive with the optimal clustering. Data variance can be used as criteria for feature selection and extraction. For example, Principal Component Analysis (PCA) is a classical feature extraction method which finds a set of mutually orthogonal basis functions that capture the directions of maximum variance in the data.

III. PROPOSED ALGORITHM

A. PRE-PROCESSING

The unsupervised raw dataset is first partitioned into three groups: (1) a finite set of objects, (2) the set of attributes (features, variables) and (3) the domain of attribute. For each groups in the dataset, a decision system is constructed. Each decision system is subsequently split into two parts: the training dataset and the testing dataset. Each training dataset uses the corresponding input features and fall into two classes: normal (+1) and abnormal (−1).

B. K-MEANS BASED PRINCIPAL COMPONENT ANALYSIS (PCA)

K-means Clustering algorithms is a widely used partitioning based technique that attempts to find a user specified number of clusters (*k*), which are represented by their centroids, by minimizing the square error function. The K-means algorithm is one of the partitioning based, non-hierarchical clustering methods. Given a set of numeric objects *X* and an integer number *k*, the K-means algorithm searches for a partition of *X* into *k* clusters that minimizes the within groups sum of squared errors. K-means based PCA is the simplest of the true eigenvector-based multivariate analyses. Regularly, its operation can be thought of as instructive the internal structure of the data in a way which best explains the variance in the data.

The following steps of the K-means based PCA algorithm are described on algorithm 1:

Algorithm 1: K-means based PCA

Step 1: *Initialization:* choose randomly *K* input data vectors to initialize the clusters.

Step 2: *Similarity Search:* for each input vector, find the cluster center that is nearest, and allocate that input vector to the corresponding cluster.

Step 3: Find the column with maximum covariance and call it as max and sort it in any order.

Step 4: *Average Update:* update the cluster centers in each group using the mean (centroid) of the input vectors assigned to that cluster

Step 5: *Ending rule:* repeat steps 2 to 4 until no more change in the value of the means.

C. ROUGH-SET BASED FEATURE SELECTION

Rough set theory is a new mathematical approach to imprecision, vagueness and uncertainty. In an information system, every object of the universe is associated with some information. Objects characterized by the same information are indiscernible with respect to the available information about them. Any set of indiscernible objects is called an elementary set. Any union of elementary sets is referred to as a crisp set- otherwise a set is rough (imprecise, vague). Vague concepts cannot be characterized in terms of information about their elements. A rough set is the approximation of a vague concept by a pair of precise concepts, called lower and upper approximations. The lower approximation is a description of the domain objects which are known with certainty to belong to the subset of interest, whereas the upper approximation is a

DOI: 10.15680/IJIRCCE.2015. 0311067

description of the objects which possibly belong to the subset. Relative to a given set of attributes, a set is rough if its lower and upper approximations are not equal.

The main advantage of rough set analysis is that it requires no additional knowledge except for the supplied data. Rough sets perform feature selection using only the granularity structure of the data. The Roughset feature selection as the process of finding a subset of features, from the original set of pattern features, optimally according to the defined criterion. Rough sets theory is based on the concept of an upper and a lower approximation of a set, the approximation space and models of sets.

An information system can be represented as,

$$S = (U, A, V, f); \qquad (3)$$

where U is the universe, a finite set of N objects $(x_1, x_2, \ldots, x_N)$ (a nonempty set), A is a finite set of attributes, $V = \mathrm{U}_{a\in A}V_a$ (where V_a is a domain of the attribute a), $f : U \times A \rightarrow V$ is the total decision function (called the information function) such that $f(x, a) \in Va$ for every $a \in A$, $x \in U$. B subset of attributes $B \subseteq Q$ defines an equivalence relation (called an indiscernibility (unnoticeable) relation) on U.

$$\mathrm{IND}(A) = \{(x, y) \in U : \text{for all } a \in B;\ f(x, a) = f(y, a) \quad \}, \qquad (4)$$

denoted also by A'. The information system can also be defined as a decision table

$$DT = (U, C \cup D, V, f), \qquad (5)$$

where C is a set of condition attributes, D is a set of decision attributes, $V = \mathrm{U}_{a\in C \cup D} V_a$, where Va is the set of the domain of an attribute $a \in Q$, $f : U \times (C \cup D) \rightarrow V$ is a total decision function (information function, decision rule in DT) such that $f(x, a) \in Vq$ for every $a \in A$ and $x \in V$. The straightforward feature selection procedures are based on an evaluation of the predictive (Entropy) power of individual features, followed by a ranking of such evaluated features and eventually the choice of the first best m features. A criterion applied to an individual feature could be either of the open-loop or closed-loop type. It can be expected that a single feature alone may have a very low predictive power, whereas when put together with others, it may demonstrate a significant predictive power.

D. ROUGH-SET BASED K- MEANS FEATURE SELECTION

The rough set base k-means feature selection process is a search process where the whole search space covers all 2^n subsets of the n features, and with each state specifying a candidate subset. A partial order could be imposed on this search space, making each child having exactly one more feature than its parents. The k-means structure of this space determines the basic issues of the heuristic feature selection process. The first step is to decide from which state in the k-search space that the search starts. We may adopt forward selection that starts with an empty feature set and successively adds features. Another approach is to employ backward elimination that starts with all features and successively removes unnecessary ones. It is also possible to start from somewhere in the middle of the search space, that is, start with a subset that contains some indispensable features and search outwards from this point. In rough set based feature selection approaches, the core can be used as the starting point. The second issue is how the search is carried out. The simplest way is the greedy method which traverses the search space without backtrack. At each step, only one feature is added or removed. Once a feature is added, it cannot be removed in later steps. Likewise, once a feature is removed, it cannot be added. Another method, known as stepwise selection or elimination, allows adding (removing) a feature that was removed (added) in the previous step. In our rough sets based k-means feature selection; we adopt the forward selection approach since all the features in cannot be removed. The thesis successively adds features until the stop criterion is satisfied. We use a measure, or heuristic function, to evaluate alternative feature subsets.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

E. GENETIC SEARCH ALGORITHM

Genetic algorithm (GA) is a search heuristic, used to generate useful solutions to optimization and search problems. In the proposed optimal rough set based genetic algorithm, the dependency of the decision feature attribute on different set of conditional variables is calculated and attributes with highest dependency value is selected as optimum reduct by applying genetic algorithm. The reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same quality of classification as the original. A reduct is defined as a subset R of the conditional attribute set C such that $\gamma_R(D) = \gamma_c(D)$. A given dataset may have many attribute reduct sets, so the set R of all reducts is defined as:

$$R = \{X: X \subseteq C, \gamma_R(D) = \gamma_c(D)\} \qquad (6)$$

The genetic search optimal reduct algorithm given in algorithm 2, attempts to calculate a minimal reduct without exhaustively generating all possible subsets. It starts with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in dependency, until this produces its maximum possible value for the dataset.

Algorithm 2: Genetic search optimal quick reduct (C, D)

C, the set of all conditional features;

D, the set of decision features.

```
Step 1: R ← {}
Step 2: do
Step 3: T ← R
Step 4: ∀x ∈ (C-R)
Step 5: if $\gamma_{R\cup\{x\}}(D) = \gamma_c(D)$
Step 6: T ← R U {x}
Step 7: R ← T
Step 8: until $\gamma_R(D) = \gamma_c(D)$
Step 9: return R
```

IV. RESULTS

The research work results describe a preliminary experimental evaluation of the feature selection and feature extraction algorithms presented in this thesis. We implemented the proposed algorithms in MATLAB and compared them against a few other prominent dimensionality reduction techniques such as the Laplacian scores [27]. Laplacian scores is a popular feature selection method for clustering and classification. We performed all the experiments on a Windows machine with a dual core 2.8 Ghz processor and 2 GB of RAM. The research work performed the experiments on a few real-world and synthetic datasets. For the synthetic dataset, we generated a dataset of m = 1000 points in n = 2000 dimensions as follows. We chose k = 5 centers uniformly at random from the n-dimensional hypercube of side length 2000 as the ground truth centers. We then generated points from a Gaussian distribution of variance one, centered at each of the real centers. To each of the 5 centers we generated 200 points (we did not include the centers in the dataset). Thus, we obtain a number of well separated Gaussians with the real centers providing a good approximation to the optimal clustering. For the real-world datasets we used five datasets that we denote by *USPS*, *COIL20*, *ORL*, *PIE* and *LIGHT*. The *USPS* digit dataset contains grayscale pictures of handwritten digits and can be downloaded from the UCI repository [29]. Each data point of USPS has 256 dimensions and there are 1100 data points per digit. The coefficients of the data points have been normalized between 0 and 1. The *COIL20* dataset contains 1400 images of 20 objects (the images of each objects were taken 5 degrees apart as the object is rotated on a turntable and each object has 72 images) and can be downloaded from [30]. The size of each image is 32×32 pixels, with 256 grey levels per pixel. To normalize data, traditionally this means to fit the data within unity, so all data values will take on a value of 0 to 1. Since some models collapse at the value of zero, sometimes an arbitrary range of say 0.1 to 0.9 is chosen instead, but for this post I will assume a unity-based normalization. The following equation is what should be used to implement a unity-based normalization:

$$X_{i,0\ to\ 1} = \frac{X_i - X_{min}}{X_{max} - X_{min}} \qquad (7)$$

Where: X_i = Each data point I, X_{Min} = The minima among all the data points, X_{Max} = The maxima among all the datapoints, $X_{i,\ 0\ to\ 1}$ = The data point i normalized between 0 and1

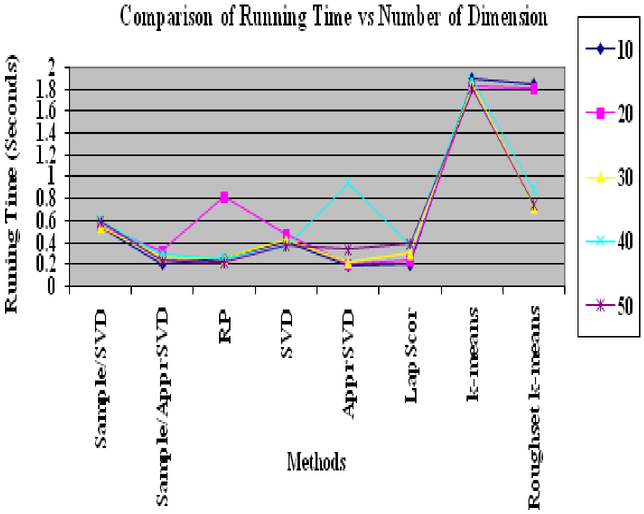

Fig 2. Running time vs Number of Dimension

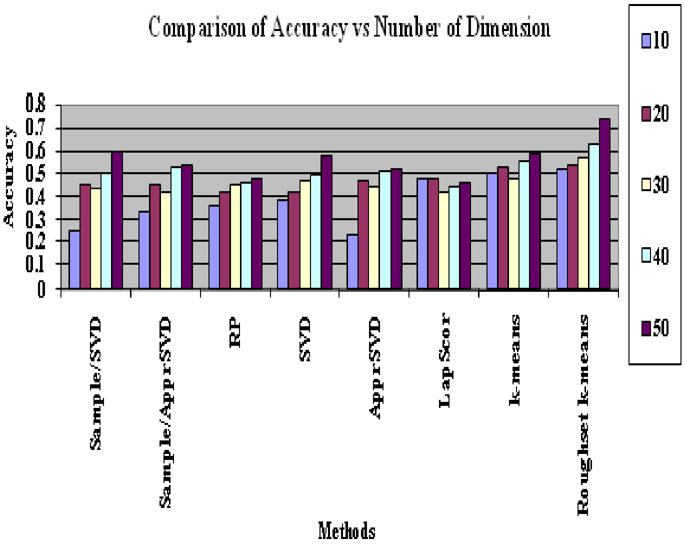

Fig 3. Accuracy vs Number of Dimension

V. CONCLUSION AND FUTURE WORK

This research presents an enhanced method such as Randomized Dimensionality Reduction using Roughset based k-means clustering which combines PCA, k-means clustering, random projections, and roughset based k-means to solve the problem of dimensionality reduction. In the introduction of this thesis tells about the overview, the objective and the contribution of the research work. Next chapter surveys the previous work related to the scope of the thesis. Then the next is the main chapter in this thesis which shows the framework of the research and the methodologies performed in this research such as PCA, k-means clustering, random projections, and roughset based k-means feature selection, genetic search algorithm.

The proposed methodologies performance is analyzed with real-world and synthetic datasets those are downloaded from UCI repository. The values are compared with several constrains such as number of dimensions versus objective, running time, accuracy. Based on the results generated this research concludes that accuracy increases compared to the previous method of k-means clustering algorithm.

A further challenge is to identify an important future direction is to develop a computationally efficient method of determining the distance metric of the embedding space, Manifold Finding and Dynamic/Streaming data. Evolving some dimensional reduction methods like canon pies can be used for high dimensional datasets is suggested as future work.

REFERENCES

1. J. A. Hartigan, Clustering Algorithms. New York, NY, USA: Wiley, 1975.
2. S. Lloyd, “Least squares quantization in PCM,” IEEE Trans. Inf. Theory, vol. 28, no. 2, pp. 129–137, Mar. 1982.
3. R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy, “The effectiveness of Lloyd-type methods for the k-means problem,” in Proc.47th Annu. IEEE Symp. Found. Comput. Sci. (FOCS), Oct. 2006, pp. 165–176.
4. X. Wu et al., “Top 10 algorithms in data mining,” Knowl. Inf. Syst., vol. 14, no. 1, pp. 1–37, 2008.

5. I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror, “Result analysis of the NIPS 2003 feature selection challenge,” in Neural Information Processing Systems. Red Hook, NY, USA: Curran & Associates Inc., 2005.
6. R. Ostrovsky and Y. Rabani, “Polynomial time approximation schemes for geometric k-clustering,” in Proc. 41st Annu. IEEE Symp. Found.Comput. Sci. (FOCS), 2000, pp. 349–358.
7. A. Kumar, Y. Sabharwal, and S. Sen, “A simple linear time (1 +)- approximation algorithm for k-means clustering in any dimensions,” in Proc. 45th Annu. IEEE Symp. Found. Comput. Sci. (FOCS), 2004, pp. 454–462.
8. S. Har-Peled and S. Mazumdar, “On coresets for k-means and k-median clustering,” in Proc. 36th Annu. ACM Symp. Theory Comput. (STOC), 2004, pp. 291–300.
9. S. Har-Peled and A. Kushal, “Smaller coresets for k-median and k-means clustering,” in Proc. 21st Annu. Symp. Comput. Geometry (SoCG), 2005, pp. 126–134.
10. G. Frahling and C. Sohler, “A fast k-means implementation using coresets,” in Proc. 22nd Annu. Symp. Comput. Geometry (SoCG), 2006, pp. 135–143.
11. D. Arthur and S. Vassilvitskii, “k-means++: The advantages of careful seeding,” in Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA), 2007, pp. 1027–1035.
12. P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, “Clustering in large graphs and matrices,” in Proc. 10th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA), 1999, pp. 291–299.
13. D. Feldman, M. Schmidt, and C. Sohler, “Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering,” in Proc. 24th Annu. ACM-SIAM SODA, 2013, pp. 1434–1453.
14. I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” J. Mach. Learn. Res., vol. 3, pp. 1157–1182, Mar. 2003.
15. W. B. Johnson and J. Lindenstrauss, “Extensions of Lipschitz mappings into a Hilbert space,” Contemp. Math., vol. 26, pp. 189–206, 1984.
16. C. Boutsidis, M. W. Mahoney, and P. Drineas, “Unsupervised feature selection for the k-means clustering problem,” in Neural Information Processing Systems. Red Hook, NY, USA: Curran & Associates Inc., 2009.
17. M. Rudelson and R. Vershynin, “Sampling from large matrices: An approach through geometric functional analysis,” J. ACM, vol. 54, no. 4,2007, Art. ID 21.
18. T. Sarlos, “Improved approximation algorithms for large matrices via random projections,” in Proc. 47th Annu. IEEE Symp. Found. Comput. Sci. (FOCS), Oct. 2006, pp. 143–152.
19. C. Boutsidis, P. Drineas, and M. Magdon-Ismail. (2011). “Near optimal column based matrix reconstruction.” [Online]. Available: http://arxiv.org/abs/1103.0995
20. M. D. Vose, “A linear algorithm for generating random numbers with a given distribution,” IEEE Trans. Softw. Eng., vol. 17, no. 9, pp. 972–975, Sep. 1991.
21. M. Magdon-Ismail. (2010). “Row sampling for matrix algorithms via a non-commutative bernstein bound.” [Online]. Available: http://arxiv.org/abs/1008.0587
22. P. Indyk and R. Motwani, “Approximate nearest neighbors: Towards removing the curse of dimensionality,” in Proc. 30th Annu. ACM Symp. Theory Comput. (STOC), 1998, pp. 604–613.
23. N. Ailon and B. Chazelle, “Approximate nearest neighbors and the fast Johnson–Lindenstrauss transform,” in Proc. 38th Annu. ACM Symp. Theory Comput. (STOC), 2006, pp. 557–563.
24. D. Achlioptas, “Database-friendly random projections: Johnson–Lindenstrauss with binary coins,” J. Comput. Syst. Sci., vol. 66, no. 4, pp. 671–687, 2003.
25. E. Liberty and S. W. Zucker, “The Mailman algorithm: A note on matrix–vector multiplication,” Inf. Process. Lett., vol. 109, no. 3,pp. 179–182, 2009.
26. MATLAB, 7.13.0.564 (R2011b), MathWorks, Natick, MA, USA, 2010.
27. X. He, D. Cai, and P. Niyogi, “Laplacian score for feature selection,” in Neural Information Processing Systems, Y. Weiss, B. Schölkopf, andJ. Platt, Eds. Red Hook, NY, USA: Curran & Associates Inc., 2006, pp. 507–514.