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 8, August 2015

A Study of Dimensionality Reduction Using Roughset Based K-Means

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: Feature Reduction of pattern dimensionality using feature extraction and feature selection belongs to the data mining. To enhance the robustness of the k-means clustering algorithm and for visualization purpose the dimension reduction techniques may be employed. Randomized Dimensionality reduction is the transformation of high-dimensional data into a significant illustration of reduced dimensionality that corresponds to the fundamental dimensionality of the data. K-means clustering algorithm often not well for high dimension datasets and error dimensionality reduction, hence, to improve the efficiency, the proposed system apply Roughset theory based k-means on original data set and obtain a reduced dataset containing possibly uncorrelated variables. In this paper, Roughset theory for feature selection and K-means based principal component analysis (PCA) for Feature Extraction, non-linear conversion is used for reduce the dimensionality and primary centroid is calculated, then it is applied to K-Means clustering algorithm.

KEYWORDS: K-means, Principal Component Analysis, Roughset, high dimension

I. INTRODUCTION

Data Mining refers to the mining or discovery of new information in terms of patterns or rules from vast amounts of data. Data mining is a process that takes data as input and outputs knowledge. Applications in various domains such as text/web mining and bioinformatics often lead to very high dimensional data. Clustering such high-dimensional data sets is a contemporary challenge, due to the curse of dimensionality. A common practice is to project the data onto a low-dimensional subspace through unsupervised dimensionality reduction such as Principal Component Analysis (PCA) and various manifold learning algorithms before the clustering.

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]. Practitioners have addressed these obstacles by introducing feature selection and feature extraction techniques. Feature selection selects a (small) subset of the actual features of the data, whereas feature extraction constructs a (small) set of artificial features based on the original features. Here, we consider a rigorous approach to feature selection and feature extraction for k-means clustering. Next, we describe the mathematical framework under which we will study such dimensionality reduction methods. Principal component analysis (PCA) [9] involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components.

DOI: 10.15680/IJIRCCE.2015. 0308039

The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

High-dimensional datasets present many mathematical challenges as well as some opportunities, and are bound to give rise to new theoretical developments [6]. One of the problems with high-dimensional datasets is that, in many cases, not all the measured variables are “important” for understanding the underlying phenomena of interest. While certain computationally expensive novel methods [4] can construct predictive models with high accuracy from high-dimensional data, it is still of interest in many applications to reduce the dimension of the original data prior to any modelling of the data.

II. RELATED WORK

In [2] authors addressed the 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 to fall. It has been shown by Panter and Dite that, in the limit as the number of quanta becomes infinite, the 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. 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. The optimization criterion used is that the average quantization noise power be a minimum. It is shown that the result obtained here goes over into the Panter and Dite result as the number of quanta become large. 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. The authors proposed and justify a cluster ability criterion for data sets. It presents variants of Lloyd's heuristic that quickly lead to provably near-optimal clustering solutions when applied to well-clusterable instances. This is the first performance guarantee for a variant of Lloyd's heuristic. The provision of a guarantee on output quality does not come at the expense of speed: some of our algorithms are candidates for being faster in practice than currently used variants of Lloyd's method. In addition, our other algorithms are faster on well-clusterable instances than recently proposed approximation algorithms, while maintaining similar guarantees on clustering quality. In [4] authors provided participants with five datasets 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. Participants were asked to make on-line submissions on the validation and test sets, with performance on the validation set being presented immediately to the participant and performance on the test set presented to the participants at the workshop. In total 1863 entries were made on the validation sets during the development period and 135 entries on all test sets for the final competition. In [5] authors discussed to deal with 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. It gives polynomial time approximation schemes for this problem in several settings, including the binary cube (0, 1)/sup d/ with Hamming distance, and R/sup d/ either with L/sup 1/ distance, or with L/sup 2/ distance, or with the square of L/sup 2/ distance. In all these settings, the best previous results were constant factor approximation guarantees. It note that our problem is similar in flavor to the k-median problem (and the related facility location problem), which has been considered in graph-theoretic and fixed dimensional geometric settings, where it becomes hard when *k* is part of the input. In [8] Authors consider the problem of dividing a set of m points in Euclidean n\Gamma space into k clusters (m; n are variable while k is fixed), so as to minimize the sum of distance squared of each point to its "cluster center". This formulation differs in two ways from the most frequently considered clustering problems in the literature, namely, here we have k fixed and m;n variable and we use the sum of squared distances as our measure; we will argue that our problem is natural in many contexts. We consider a relaxation of the discrete problem : find the k\Gamma dimensional subspace V so that the sum of distances squared to V (of the m points) is minimized. Its shows: (i) The relaxation can be solved by Singular Value Decomposition (SVD) of Linear Algebra. (ii) The solution of the relaxation can be used to get a 2-approximation algorithm for the original problem.

DOI: 10.15680/IJIRCCE.2015. 0308039

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)Algorithm:*

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 centre 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 centres 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:*

The Roughest 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,

$$S = (U, A, V, f); \quad (1)$$

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, $V = U_{a \in A} V_a$ (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.

IV. Conclusion and Future Work

In this paper a dimensionality reduction through Roughset based K-means Clustering algorithm. Using randomized dimension reduction of roughest theory, original real-world and synthetic datasets is compact to reduced data set which was partitioned in to k clusters in such a way that the amount of the total clustering errors for all clusters was reduced as much as possible while inter distances between clusters are maintained to be as large as possible. The proposed algorithm is to initialize the clusters which are then applied to k-means algorithm. Developing some new dimensional reduction methods like canon pies can be used for high dimensional datasets is suggested as future work.

References

1. Christos Boutsidis, Anastasios Zouzias, Michael W. Mahoney, and Petros Drineas, "Randomised Dimensionality Reduction for K-Means Clustering", IEEE transactions on information vol. 61, no. 2, February 2015.
2. J. A. Hartigan, "Clustering Algorithms". New York, NY, USA: Wiley, 1975.
3. S. Lloyd, "Least squares quantization in PCM," IEEE Trans. Inf. Theory, vol. 28, no. 2, pp. 129–137, Mar. 1982.

4. 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.
5. X. Wu et al., "Top 10 algorithms in data mining," Knowl. Inf. Syst., vol. 14, no. 1, pp. 1–37, 2008.
6. 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.
7. D.L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. Lecture delivered at the "Mathematical Challenges of the 21st Century" conference of The American Math. Society, Los Angeles, August 6-11.
8. L. Breiman. Random forests. Technical report, Department of Statistics, University of California, 2001.
9. 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.
10. 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.