International Journal of Enhanced Research in Management & Computer Applications, ISSN: 2319-7471
Vol. 3 Issue 1, January-2014, pp: (7-14), Impact Factor: 1.147, Available online at: www.erpublications.com

A Dimension Reduction for Cluster Analysis using ODC and SODC

Shailesh Singh Panwar[1], Lokesh Singh Panwar[2]
Department of Computer Science and Engineering, Graphic Era University Dehradun, India
Department of Instrumentation and and Engineering, HNB Garhwal University, Srinagar Garhwal, India

Abstract: Drawing scatter plots is usually the first step when conducting a cluster analysis. For high-dimensional data, scatter plots are usually drawn on the first few principal components. However, principal component analysis doesn't take into account the clustering structure and therefore such scatter plots may be misleading. In this manuscript, we reinvestigate an existing method, optimal discriminant clustering and propose to use it as a dimension reduction tool for cluster analysis. Furthermore, because in high-dimensional data many of the features may be non- informative for clustering, we propose sparse optimal discriminant clustering (SODC) by adding a variation of group-lasso penalty to ODC.

Keywords: Data Reduction, Dimension Data Reduction, Clusting, ODC, SODC.

I. INTRODUCTION

With the tremendous growth of computer networks mostly computer system suffers from security vulnerabilities which are difficult to handle technically as well as economically by users [1].A process in which amount of data is minimized and that minimized data are stored in a data storage environment is known as data reduction. By data reduction algorithms reduce massive data-set to a manageable size without significant loss of information represented by the original data. This process various advantages have achieved in computer networks such as increasing storage efficiency and reduce costs. There are two important motivating factors of data reduction, first is redundancy and second is reduction of complexity regarding live network acquisition [2].

Data reduction technique can be applied to obtain a reduce representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on reduced data set should be more efficient yet produce the same (or almost the same) analytical results. For very large data sets, there is an increased likelihood that intermediate, additional steps, data reduction, should be performed prior to applying the data reduction techniques. Three basic operations in data reduction process are: delete a column, delete a row, and reduce the number of column.

The larger data sets have more useful information but it decreasing storage efficiency and increasing costs but larger data set have more useful information. Which techniques, methods or relative terms are used in larger data set may also used in smaller data set. So the benefit of data reduction techniques we propose increase as the data sets themselves increase size, complexity and reduce the costs. We have taken a broad view of large qualitative data sets, aiming to highlight trends, relationships, or associations for further analysis, without loss of any information [3].

The recent advantage data collection and storage capabilities have include information overhead in many applications sciences, e.g., on-line monitoring of spacecraft operations with time series data. In this we perform data reduction technique before storage or transmission of data. This can be some information loss, but not all features of data might be relevant. The motivating factor of this is that real system, which data we get after data reduction, that dimensionality is lower than the space, which is measure in. Reconstruct the lower dimensional samples which are needed and possible with varying degrees of accuracy [4].

II. DATA REDUCTION

Data reduction is the transformation of numerical or alphabetical digital information derived empirical or experimentally into a corrected, ordered, and simplified form. The basic concept is the reduction of multitudinous amounts of data down to the meaningful parts. By data reduction reduce massive data-set to a manageable size without significant loss of information represented by the original data.

The advantages of data reduction are results are shown in a compact form and easy to understand. The graphical or pictorial representations can be used. Overall patterns can be seen. In this comparisons can be made between different sets of data. The quantitative measures can be used. The disadvantages are original data are lost and the process is irreversible.

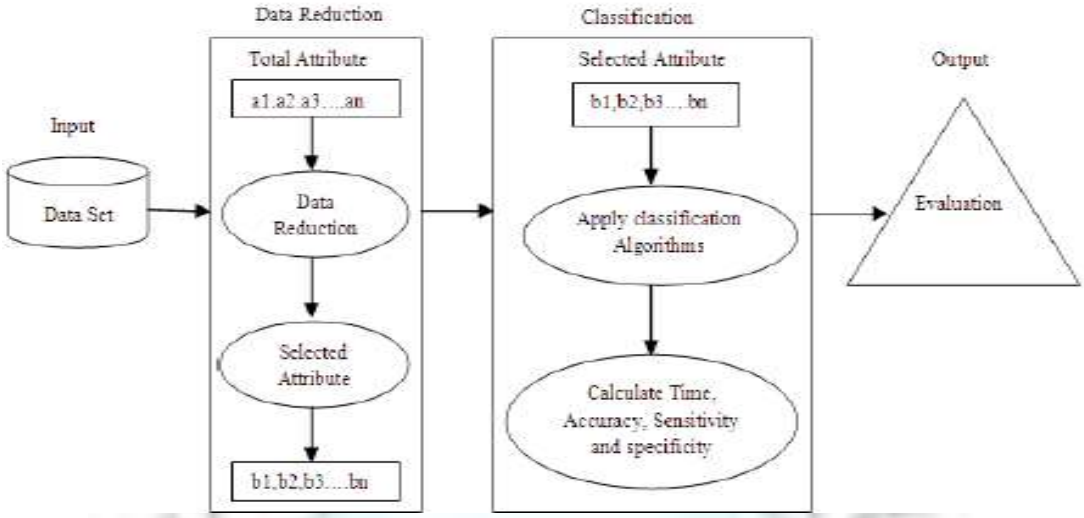

Figure 1 Data Reduction and Classification Process

Data Reduction Scope

Data Reduction Reduce the size of massive data-set to a manageable size without significant loss of information represented by the original data and also reduces the communications costs and decrease storage requirements. Data reduction also has some more scopes.

First is Primary Storage which reduces physical capacity for storage of active data. Second is Replication, reduce capacity for disaster recovery and business continuity. Third one is Data Production; reduce capacity for backup with longer retention periods. Fourth is Archive, reduce capacity for retention and preservation. Fifth is Movement/Migration of data, reduce bandwidth requirements for data-in-transit [5].

The Advantageof data reduction is the results are shown in a compact form and easy to understand. The graphical or pictorial representations can be used. Overall patterns can be seen. In this comparisons can be made between different sets of data. The quantitative measures can be used [6], and the Disadvantage ofdata reduction is details of the original data are lost and the process is irreversible [6].

The Limitation of data reduction is, Database systems are complex, difficult, and time-consuming to design. It has substantial hardware and software start-up costs. Damage to database affects virtually all applications programs. Extensive conversion costs in moving form a file-based system to a database system. In this initial training required for all programmers and users [7].

There are three data reduction strategies:

1. Dimensionality Reduction: - Dimensionality Reduction is about converting data of very high dimensionality into data of much lower dimensionality such that each of the lower dimensions conveys much more information [8]. Feature selection (i.e., attribute subset selection) is selecting a minimum set of attributes (features) that is sufficient for the data mining task. Heuristic methods is step-wise forward selection and step-wise backward elimination. It is combining forward selection and backward elimination [9].

It is so easy and convenient to collect data an experiment. Data is not collected only for data mining. Data accumulates in an unprecedented speed. Data preprocessing is an important part for effective machine learning and data mining. Dimensionality reduction is an effective approach to downsizing data. Most machine learning and data mining techniques may not be effective for high-dimensional data. Dimensionality Reduction works in three ways, first is Visualization,

means projection of high-dimensional data onto 2D or 3D. Next is Data compression means efficient storage and retrieval and then Noise removal means positive effect on query accuracy [5].

2. Clustering:- Clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics [10].Clustering is partition data set into clusters, and one can store cluster representation only. It can be very effective if data is clustered but not if data is "smeared". There are many choices of clustering definitions and clustering algorithms

3. Sampling: - It is used in conjunction with skewed data. Sampling obtaining a small sample to represent the whole data set. It Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data. Key principle of sampling chooses a representative subset of the data. Simple random sampling may have poor performance in the presence of skew. It develops adaptive sampling methods. Stratified sampling is approximate the percentage ofeach class (or subpopulation of interest) in the overall database [11].

III. METHODOLOGY

Optimal Discriminant Clustering (ODC)

An n*(k-1) matrix Y is called the sample scoring matrix if it satisfies [12]

$$YY' = I_{(k-1)} and 1_n' Y = 0$$

The main part of ODC is a minimization process,

$$(\hat{W}, \hat{Y}) = \arg\min_{W,Y} \|Y - H_n XW\|_F^2 + \lambda_2 \|W\|_F^2, s.t, YY' = I_{(k-1)} and 1_n' Y = 0 \quad (1)$$

Where W is a p*(k-1) matrix and

$$\|A\|_F = \sqrt{tr(AA')}$$

is the Frobenius norm of any matrix. Hn is centering matrix, and In is Identity matrix, 1n is the n-dimensional vector of ones.
Optimal Discriminant Clustering (ODC) algorithm:

Step 1: Obtain W from (1) given µ2.
Step 2: Calculate Z=(Z1…Zn) = HnXW.
Step 3: Perform k-means on Zi, I =1…n.
Step 4: return the partition of Zi as the partition of Xi.

Sparse Optimal Discriminant Clustering (SODC)

It is hard to interpret when it involves all the features. Moreover, in situations where many features are not informative for clustering, including such features may cause poor performance of ODC[13]. Therefore, we propose to add group-lasso type of penalty in (1) to obtain sparse solution of W.

$$(\hat{W}, \hat{Y}) = \arg\min_{W,Y} \|Y - H_n XW\|_F^2 + \lambda_2 \|W\|_F^2 + \lambda_1 \sum_{j=1}^{p} \|w_j\|_2, s.t, YY' = I_{(k-1)} and 1_n' Y = 0 \quad (2)$$

Use ODC to get initial value of Y and W.Given W, obtain Y using Y=UV, where Hn XW=UDV.Given Y[14], obtain W using block coordinate algorithm:-Use same step 2,3,4 as in ODC .
Rewrite Z as:

$$(z^{(1)}, \ldots, z^{(k-1)})$$

And call the jth SODC component.

IV. SIMULATION STUDY

Dataset Generation

Generate data consists of 100 observations with Variables. The first informative variables are generated from [15]

$$N(m(C_i),I_q)$$

Where q is even and

$$m(C_i)=u(-1_{q/2},1_{q/2})'I(C_i=1)+u1_q I(C_i=2)+u(1_{q/2},-1_{q/2})'I(C_i=3);$$

The three clusters are well separated when u is large, and can be heavily overlapped when u is small[16]. The last p-q noise variables are generated from:-

$$N(0_{p-q},1_{p-q})$$

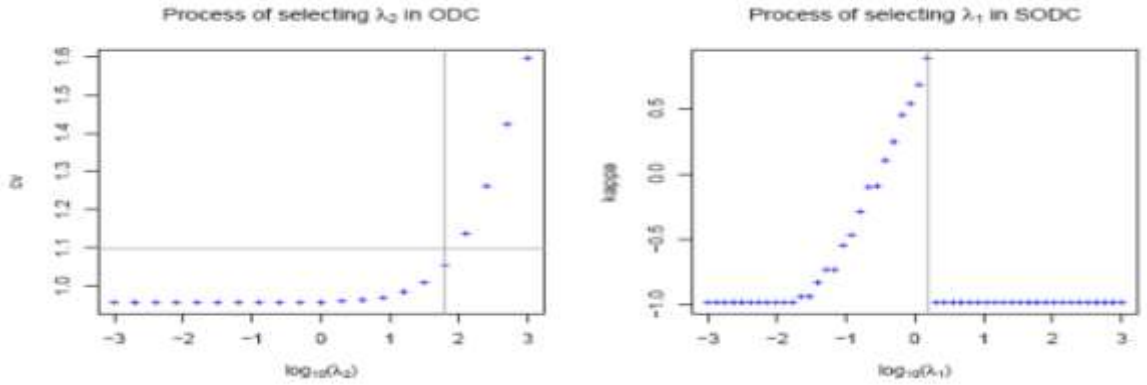

Figure 2 Choose Tuning Parameter

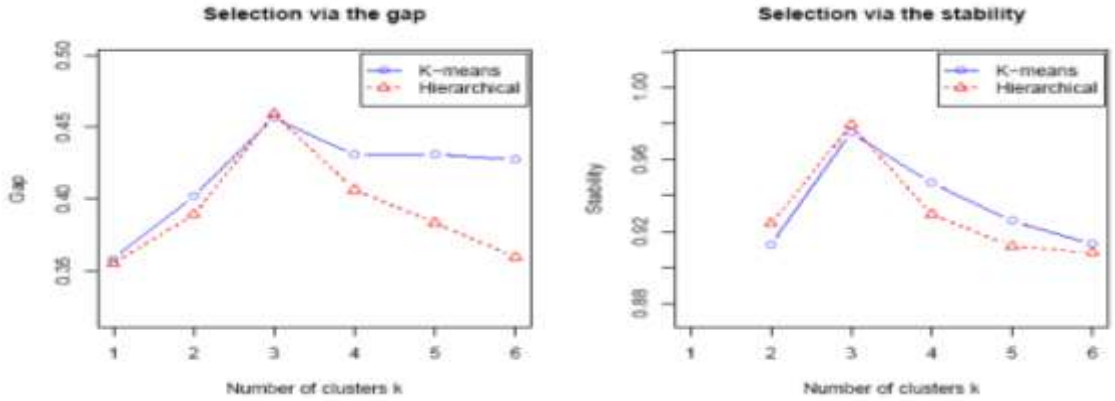

Figure 3 Selection of number of Clusters

Table 1 Selection of number of Clusters k via the Gap Statistic

p	μ	Distribution of $\hat{k}$					
		1	2	3	4	5	6+
10	2.0	0	0	49	0	1	0
	2.2	0	0	45	1	0	4
	2.4	0	0	48	1	1	0
50	2.0	0	0	50	0	0	0
	2.2	0	0	50	0	0	0
	2.4	0	0	50	0	0	0
100	2.0	0	12	38	0	0	0
	2.2	0	0	50	0	0	0
	2.4	0	0	50	0	0	0
200	2.0	27	23	0	0	0	0
	2.2	7	34	9	0	0	0
	2.4	0	7	43	0	0	0

Table 2: Comparing six clustering procedures: (1) using only informative features; (2) using all features; (3) using some PCA components; (4) using the first two ODC components; (5) using the first two SODC components; and (6) using only the features selected by SODC.

		$\mu = 2$		$\mu = 2.2$		$\mu = 2.4$	
p	Method	ARI (SD)	CE (SD)	ARI (SD)	CE (SD)	ARI (SD)	CE (SD)
10	ORACLE	0.90(0.05)	0.02(0.01)	0.93(0.04)	0.01(0.01)	0.96(0.03)	0.01(0.01)
	ALL	0.90(0.05)	0.02(0.01)	0.93(0.04)	0.02(0.01)	0.96(0.03)	0.01(0.01)
	PCA	0.70(0.09)	0.07(0.02)	0.73(0.09)	0.06(0.02)	0.82(0.06)	0.04(0.01)
	ODC	0.85(0.05)	0.03(0.01)	0.91(0.04)	0.02(0.01)	0.94(0.03)	0.01(0.01)
	SODC	0.88(0.06)	0.03(0.01)	0.93(0.05)	0.02(0.01)	0.95(0.03)	0.01(0.01)
	SODC*	0.90(0.05)	0.02(0.01)	0.93(0.04)	0.01(0.01)	0.96(0.03)	0.01(0.01)
50	ORACLE	0.91(0.03)	0.02(0.01)	0.95(0.03)	0.01(0.01)	0.96(0.03)	0.01(0.01)
	ALL	0.88(0.03)	0.03(0.01)	0.92(0.03)	0.02(0.01)	0.96(0.03)	0.01(0.01)
	PCA	0.64(0.10)	0.08(0.02)	0.72(0.08)	0.06(0.02)	0.79(0.06)	0.05(0.01)
	ODC	0.82(0.05)	0.04(0.01)	0.87(0.05)	0.03(0.01)	0.94(0.03)	0.01(0.01)
	SODC	0.86(0.08)	0.03(0.02)	0.93(0.04)	0.02(0.01)	0.96(0.03)	0.01(0.01)
	SODC*	0.91(0.03)	0.02(0.01)	0.95(0.03)	0.01(0.01)	0.96(0.03)	0.01(0.01)
200	ORACLE	0.92(0.05)	0.02(0.01)	0.94(0.02)	0.01(0.00)	0.96(0.03)	0.01(0.01)
	ALL	0.78(0.07)	0.05(0.02)	0.86(0.04)	0.03(0.01)	0.90(0.04)	0.02(0.01)
	PCA	0.59(0.11)	0.09(0.02)	0.67(0.07)	0.07(0.02)	0.75(0.08)	0.05(0.02)
	ODC	0.65(0.11)	0.08(0.02)	0.78(0.06)	0.05(0.01)	0.82(0.08)	0.04(0.02)
	SODC	0.84(0.07)	0.04(0.02)	0.92(0.03)	0.02(0.01)	0.93(0.04)	0.02(0.01)
	SODC*	0.92(0.05)	0.02(0.01)	0.94(0.02)	0.01(0.00)	0.96(0.03)	0.01(0.01)

Table 2: Comparing six clustering procedures: (1) using only informative features; (2) using all features; (3) using some PCA components; (4) using the first two ODC components; (5) using the first two SODC components; and (6) using only the features selected by SODC.

		$\mu = 2$		$\mu = 2.2$		$\mu = 2.4$	
p	Method	ARI (SD)	CE (SD)	ARI (SD)	CE (SD)	ARI (SD)	CE (SD)
10	ORACLE	0.90(0.05)	0.02(0.01)	0.93(0.04)	0.01(0.01)	0.96(0.03)	0.01(0.01)
	ALL	0.90(0.05)	0.02(0.01)	0.93(0.04)	0.02(0.01)	0.96(0.03)	0.01(0.01)
	PCA	0.70(0.09)	0.07(0.02)	0.73(0.09)	0.06(0.02)	0.82(0.06)	0.04(0.01)
	ODC	0.85(0.05)	0.03(0.01)	0.91(0.04)	0.02(0.01)	0.94(0.03)	0.01(0.01)
	SODC	0.88(0.06)	0.03(0.01)	0.93(0.05)	0.02(0.01)	0.95(0.03)	0.01(0.01)
	SODC*	0.90(0.05)	0.02(0.01)	0.93(0.04)	0.01(0.01)	0.96(0.03)	0.01(0.01)
50	ORACLE	0.91(0.03)	0.02(0.01)	0.95(0.03)	0.01(0.01)	0.96(0.03)	0.01(0.01)
	ALL	0.88(0.03)	0.03(0.01)	0.92(0.03)	0.02(0.01)	0.96(0.03)	0.01(0.01)
	PCA	0.64(0.10)	0.08(0.02)	0.72(0.08)	0.06(0.02)	0.79(0.06)	0.05(0.01)
	ODC	0.82(0.05)	0.04(0.01)	0.87(0.05)	0.03(0.01)	0.94(0.03)	0.01(0.01)
	SODC	0.86(0.08)	0.03(0.02)	0.93(0.04)	0.02(0.01)	0.96(0.03)	0.01(0.01)
	SODC*	0.91(0.03)	0.02(0.01)	0.95(0.03)	0.01(0.01)	0.96(0.03)	0.01(0.01)
200	ORACLE	0.92(0.05)	0.02(0.01)	0.94(0.02)	0.01(0.00)	0.96(0.03)	0.01(0.01)
	ALL	0.78(0.07)	0.05(0.02)	0.86(0.04)	0.03(0.01)	0.90(0.04)	0.02(0.01)
	PCA	0.59(0.11)	0.09(0.02)	0.67(0.07)	0.07(0.02)	0.75(0.08)	0.05(0.02)
	ODC	0.65(0.11)	0.08(0.02)	0.78(0.06)	0.05(0.01)	0.82(0.08)	0.04(0.02)
	SODC	0.84(0.07)	0.04(0.02)	0.92(0.03)	0.02(0.01)	0.93(0.04)	0.02(0.01)
	SODC*	0.92(0.05)	0.02(0.01)	0.94(0.02)	0.01(0.00)	0.96(0.03)	0.01(0.01)

Table 3: Feature selection by SODC, compared with Raftery and Dean's model-based clustering with headlong algorithm (MCLH) and Witten and Tibshirani's sparse k-means clustering (SKM). True: the percentage of selecting exactly the true subset of informative features; FP: the average number of incorrectly selected noise variables; FN: the average number of incorrectly excluded informative features; Size: the average size of the selected subset.

		$\mu = 1.2$			$\mu = 1.4$			$\mu = 1.6$		
p	Method	True	FP/FN	Size	True	FP/FN	Size	True	FP/FN	Size
10	SODC	85%	0.15/0.00	2.15	95%	0.05/0.00	2.05	100%	0.00/0.00	2.00
	SKM	65%	2.05/0.00	4.05	100%	0.00/0.00	2.00	100%	0.00/0.00	2.00
	MCLH	70%	0.10/0.20	1.90	100%	0.00/0.00	2.00	100%	0.00/0.00	2.00
50	SODC	75%	0.30/0.00	2.30	95%	0.05/0.00	2.05	100%	0.00/0.00	2.00
	SKM	35%	3.40/0.00	5.40	100%	0.00/0.00	2.00	95%	0.05/0.00	2.05
	MCLH	65%	0.70/0.25	2.45	90%	0.15/0.00	2.15	90%	0.15/0.00	2.15
200	SODC	0%	3.15/0.10	5.05	70%	0.30/0.05	2.25	100%	0.00/0.00	2.00
	SKM	25%	6.70/0.00	8.70	70%	2.00/0.00	4.00	90%	0.40/0.00	2.40
	MCLH	25%	4.75/0.70	6.05	30%	1.55/0.00	3.55	35%	1.25/0.00	3.25

V. ANALYSIS OF RESULTS

Table 1: It seems the gap statistic works very well when p = 10, 50, and 100. But it doesn't work well when p = 200 and = 2.0 and 2.2, which also indicates the necessity of conducting variable selection in cluster analysis.

Table 2, we see that, as a dimensional-reduction tool, both ODC and SODC perform much better than PCA, which is not working well when p is large. We also see that SODC improves ODC a lot when p is large. Moreover, we find that SODC performs similarly with ORACLE, meaning that SODC performs well in terms of feature selection. Note that sometimes ODC performs worse than the regular k-means including all the features. This fact is actually the motivation for proposing SODC. In addition, it seems that SODC always outperforms SODC, because when SODC shrinks those non-informative effects to zero, it also introduces estimation bias to those informative effects.

Table 3, we see that these three methods are comparable and very efficient in selecting the true subset including only the informative features when p is moderately large such as p = 10 or 50. It is still efficient for SODC and Witten and Tibshirani's method when = 1.4 or 1.6 and p is large such as 200. We also see that the false negative rate is very small in all of the cases and the false positive rate is also small in most of the cases.

VI. REAL DATA STUDY

Real Dataset

Table 4: Some information of the UCI datasets, where k is the number of classes (treated as clusters here), p is the number of features, and n is sample size.

Dataset	k	p	n
Wine	3	13	178
WBC	2	30	569
LM	15	90	360
DERM	6	33	366
SHD	10	256	1593

Wine Data Simulation Result

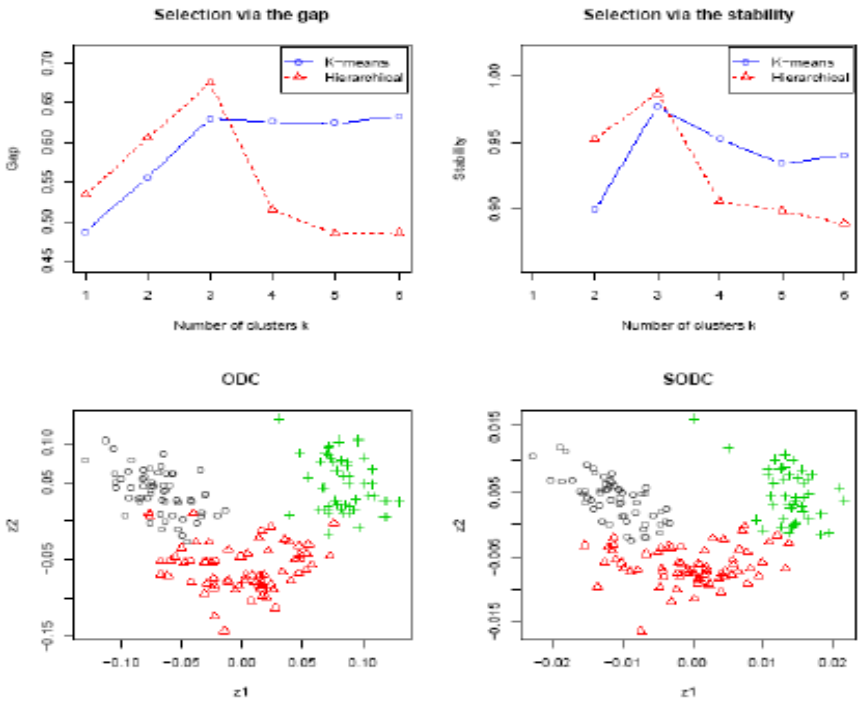

Figure 4 Wine data.

The selection processes of selecting k via the gap statistics and the stability are shown in the top panel; both k-means and hierarchical clustering are applied. Scatterplots based on first two ODC components and first two SODC components are

Clustering Method Comparison on UCI Data

Dataset	Method	ARI	CE	$\widehat{q}$
Wine	k-means	0.90	0.02	13
	ODC	0.91	0.02	13
	SODC	0.83	0.04	7
	SODC*	0.85	0.03	7
WBC	k-means	0.68	0.08	30
	ODC	0.66	0.08	30
	SODC	0.65	0.09	16
	SODC*	0.66	0.08	16
LM	k-means	0.32	0.05	90
	ODC	0.35	0.05	90
	SODC	0.25	0.05	36
	SODC*	0.30	0.05	36
DERM	k-means	0.71	0.05	33
	ODC	0.71	0.05	33
	SODC	0.72	0.04	23
	SODC*	0.73	0.04	23
SHD	k-means	0.35	0.06	256
	ODC	0.28	0.07	256
	SODC	0.31	0.07	146
	SODC*	0.33	0.07	146

VII. CONCLUSIONS

Here we reinvestigate an existing method, ODC, and advocate it as a dimension reduction tool for cluster analysis. We propose a cross-validation method for selecting the tuning parameter in ODC. We also examine the performance of using existing methods such as the gap statistic and the stability selection to select the number of clusters in ODC. As a dimension reduction tool for cluster analysis, ODC performs much better than PCA, which does not take into account the clustering structure. Furthermore, we propose SODC by adding a group-lasso type of penalty on ODC to conduct cluster analysis and feature selection simultaneously.
Both ODC and SDOC can be used as a dimension reduction tool for data visualization in cluster analysis.

REFERENCES

[1]. P. Tao, C. Xiaosu, L. Huiyu and C. Kai, "'Data Reduction for Network Forensics Using Manifold Learning", Sch. of Computer Sci. & Technol., Huazhong Univ. of Sci. & Technology . Wuhan, China, pp. 1-5, 2010.
[2]. M. Rouse, "Data Reduction", Last Updated on August 10, [Available Online] http://searchdatabackup.techtarget.com/definition/data- reduction.
[3]. E. Namey, G. guest, L. Thairu, L. Johnson, "Data Reduction Techniques for Large Qualitative Data Sets", 2007, pp 137- 162 [Available Online] http://www.stanford.edu/~thairu/07_184.Guest.1sts.pdf..
[4]. R. Georgescu, C. R. Berger, P. Willett, M. Azam _, and S. Ghoshal ," Comparison of Data Reduction Techniques Based on the Performance of SVM-type Classifiers". Dept. of Electr. and Comp. Engineering, University of Connecticut, Storrs, CT 06269, Qualtech Systems Inc., Wetherseld, USA , 2010.
[5]. T. Rivera "Advanced Data Reduction Concepts – SNIA", [Available Online] http://www.snia.org/sites/default/education/tutorials/2012/spring/data/GeneNagle_ThomasRivera_Advanced_Data_Reduction_Concepts_2.pdf.
[6]. Advantages and disadvantages of data reduction [Available Online]http://www.cl500.net/pros_cons.html.
[7]. "Limitations of the data reduction", [Available Online], www. star.bris.ac.uk /~mbt/docs / ssn69.htx/node9.html.
[8]. A. Ghodsi, "Dimensionality Reduction", Technical Report, 2006-14, Department of Statistics and Actuarial Science, University of Waterloo, pp. 5-6, 2006 .
[9]. Ricardo Gutierrez, "Dimensionality reduction", Lecture Notes on Intelligent Sensor Systems, Wright State University [Available Online] http://research.cs.tamu.edu/prism/lectures/iss/iss_l10.pdf .
[10]. Cluster Analysis, [Available Online], http://en.wikipedia.org/wiki/Cluster_analysis.
[11]. Data Preprocessing, [AvailableOnline],www.cs.uiuc.edu/homes/hanj/cs412/bk3/03 Preprocessing.ppt.
[12]. Zhang,"Optimal Scoring for Unsupervised Learning, "Advances in Neural Information . Processing Systems" 23, 12, 2241-2249, 2009.
[13]. R.Tibshirani, "Estimating the Number of Clusters in a Data Set via the Gap Statistic, "Journal of Royal Statistical Society, Series B, 63(2), 411- 423, 2001.
[14]. W.Sun,, "Consistent Selection of Tuning Parameters via Variable Selection stability, "arXiv:1208.3380v1, 2012.
[15]. A. E. Raftery.,"Variable Selection for Model-Based Clustering," Journal of the American Statistical Association, 101(473), 168-178, 2006.
[16]. Y.Fang, "Selection of the Number of Clusters via the Bootstrap Method, "Computational Statistics and Data Analysis, 56(3), 468-477, 2012.