Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 15 (2): 271-276, 2015
ISSN 1812-5654

ansinet
Asian Network for Scientific Information

RESEARCH ARTICLE OPEN ACCESS

DOI: 10.3923/jas.2015.271.276

Compression of Market Research Data Using Clustering

[1]Biku Abraham, [2]Varghese Paul and [3]Nebu John Abraham
[1]Department of Computer Applications, Saintgits College of Engineering, Pathamuttom, Kottayam, Kerala, India
[2]Department of Computer Science and Information Technology, Toc H Institute of Science and Technology, Keralal, India
[3]Malayala Manorama

ARTICLE INFO
Article History:
Received: July 10, 2014
Accepted: November 15, 2014

Corresponding Author:
Biku Abraham,
Department of Computer Applications,
Saintgits College of Engineering,
Pathamuttom, Kottayam, Kerala, India
Tel: 91 94477 97909

ABSTRACT
The emergence of consumer sovereignty has its direct impact on the collection of qualitative and quantitative data for various business purposes. Some of the data set contains comprehensive data relating to household and national level. Identification of useful dimensions underlying the data is therefore crucial. Traditionally clustering is one of the most widely used techniques for data reduction to detect common characteristics for marketing decisions. However this is only one of its uses. What is little known is its application in compressing data to save limited server space. Clustering method can also be efficiently used to encrypt and decrypt the behavioural, categorical and ratio data on some common attributes. This study contributes to the usage of cluster centroid based compression and indicates its use of centroid based dictionary coding for future decompression.

Key words: Compression, cluster, centroids, correlation, k-means, hierarchical

INTRODUCTION

Market research is widely used to collect behavioural, ratio and categorical data to quantify the general positions that a consumer have on the wide variety of product and the services. This data often helps the marketing for strategic decisions. Behavioural data is defined as the information used in marketing or industry for designing promotional campaigns based on consumer buying habits, brand preferences and product usage. To capture behavioural data, market researchers often used various scaling techniques ranging from 1-7 or 1-10 depending upon the complexity of data. It uses categorical and ratio scale variables also for the purpose. Categorical scale is of nominal type of characteristics such as gender whereas ratio scale is continuous scale example height, age. Clustering is one of the data reduction techniques which are used commonly for the approximation of behavioural segments of various consumers. In market research, clustering method is generally used to reduce the number of clusters as much as possible since it helps to manage the various associated segments in a useful way. Thus the outcome may result in an approximation of the behavioural positions of different segments. However, this ignores the potential power of clustering method to compress and to decompress large set of data inclusive of various dimensions at a later stage. The present study departs from the traditional method of minimizing the clusters and instead follows the method of increasing the clusters in a desirable limit to efficiently encrypt the original sense and soul of the data. The data set thus encrypted may be highly useful to decrypt so that the original data set can be retrieved without too much loss from the general behaviour which is applicable for marketing decisions. A detailed study on the topic may be looked up on to see as how and to what extent a clustering method has been used with an objective of compressing and decompressing the behavioural data.

Statistics mainly help the researchers to extract information (Lambert, 2003) from huge data sets and to evaluate the performance of the methods. Keeping huge data sets in memory is expensive. Different data reduction techniques are available. Cluster is one of the data reduction techniques. Clustering methods (Zhang *et al*., 2006) has been advanced in the recent years to include categorical data in addition to the ratio variables. It sequentially extracts clusters from the dataset and compared with the K modes and

Auto class. Since the inclusion of categorical data is a major breakthrough, the market researchers can look into the patterns of various segments with more accuracy. There were many studies on clustering methods. Clustering method was found good in compressing music based on string compression (Cilibrasi *et al*., 2004). The authors find that the same method can be used for data mining to identify unknown patterns hidden in the data set. The work uses similarity metrics for calculating the similarities between pieces. The significance of using clustering method to find the homogeneity (Sabater *et al*., 2004) of regional areas to distinguish urban and rural canters has been examined earlier. In a more interesting study on information (Slonim *et al*., 2005) based clustering, the researcher stated the possibility of clustering based on collective notions rather than pair wise comparisons. There is study on the speech compression (Smith, 1995) which provides telephone directory assistance and helps reducing the service time of customers. Medical image clustering (Perlmutter *et al*., 1998) study discovered a method using clustering and coding to compress digital images without much loss. In the medical context, the purpose is here to detect rather than to get the accurate information. There is an attempt to cluster the objects on the basis of subsets of attributes (Friedman and Meulman, 2004). In a thought provoking research by Cerra and Datcu (2012) suggested a fast compression method using similarity measures which could retrieve images using FCD approach.

From the review of literature it can be seen that the method of clustering and its potential usage of decompression of behavioural data has been ignored and under researched. In the present study we attempt to examine the feasibility of this method to compress the market research data used commonly from the primary and secondary sources. The study also attempts to verify the question of its closeness to the original data. Thus it is basically related to the information theory. Information is defined as: (1) Knowledge derived by study, experience or instruction, (2) Knowledge of a specific event or situation, intelligence, (3) A collection of facts or data and (4) The act of informing or the condition of being informed, communication of knowledge.

To summarize, the objectives of the study are to explore the use of clustering technique for compressing market research data and to find out the similarity of original data with the data that is replaced with cluster centroids.

MATERIALS AND METHODS

This study differs from traditional approach and tries to perceive it from the angle of using clustering method in compressing the data for saving the disk space which is also retrievable at a later stage depending upon the need from the cluster centroids generated for each data set. This calls for observation and experimentation of various issues related to this to prove the cluster centroid is close to original data. The data is collected from a mail questionnaire from exclusive showrooms and dealers of cars. The data include 126 brands of cars on four variables. The data is measured in ratio scale. This study therefore, uses market research data in the automobile sector. The study is performed in three different stages.

In the first stage the objective is to find out how to arrive at the cluster centroids which is very close to the original data. Having this purpose in the agenda, a comprehensive data collected for different branded cars and their unique characteristics mostly ratio scale is adopted. In order to find the appropriate cluster centroids, we have been increased the number of clusters from two to seven. Corresponding to each number of clusters the standard deviation of differences between cluster centroid and the original data were recorded for each observation. The equations used for finding standard deviations of difference of each observation from original data and cluster centroid are given as follows:

$$\sigma^* = \sqrt{\sum x^2 / N} \quad (1)$$

where, x = OD-CC, OD is the original data and CC is the cluster centroid. x = (x-X) and N is the total number of observations, X is the actual mean of the series.

In the second stage the aim is to find out whether the relationship remains stable with more variables. For this, the variables are increased from two to four and the corresponding results are recorded.

In the final stage, the correlation and regression techniques have been used to explore the exact relationship between the number of clusters and standard deviation. A paired t-test was performed to check the similarity between the pre clustered and post clustered data. The number of clusters were fixed using K means method and 2 stage clustering. The most popular clustering algorithm used in scientific and industrial applications is the K means algorithm which is a partitioning algorithm.

Hypothesis:

- There is an inverse relationship with standard deviation within and number of clusters
- The inverse relationship between standard deviation and number of clusters is linear in nature
- There is no difference between original data and data that is replaced with cluster centroids

RESULTS

This study aims to explore the use of clustering in market research data for compression. The clustering technique was effectively used in music notations, images etc. As pointed out earlier comparison of standard deviation and clusters has been performed. Comparison of standard deviation and clusters for two variables i.e., price and engine is given in Table 1.

Standard deviation of every cluster and its centroids for price and engine for two clusters may be taken to start with. The cluster centroid for two clusters is 36.755 and its

corresponding standard deviation from the original data is 12.2. Similarly the cluster centroid of engine is 3.675 and its corresponding standard deviation is 0.9. The standard deviations of price and engine decreased with every increase in the number of clusters up to seven. This initial result of a potential negative relationship between the number of clusters and standard deviations indicates the closeness of cluster centroid and original data with a meaningful large number of clusters with behavioural or characteristic data. To further confirm the relationship trials are done for the same data set with additional variable horsepower. The results of the standard deviation and number of clusters for three variables can be seen in Table 2.

The cluster centroid for price is 28.05, engine is 3.11 and horsepower is 188.8 for two cluster case. The corresponding standard deviations are 10.1, 0.7 and 36.7. As the number of clusters are increased the average standard deviation in each clusters (the differences between the cluster centroids and the original data) decrease as in the case of two variable experiments. A four variable comparison of similar data has been attempted to find out the stability of the finding with one more variable viz., miles per gallon. The outcome using four variables is given in Table 3.

The standard deviations of two clusters start with the 10.6 for the price, 0.7 for the engine, 37.4 for horse power and 3.3 for mpg. It is observed that the standard deviation tend to decrease with more number of clusters. Thus it is noticed that there is a relationship between standard deviation of the differences between cluster centroids and original data for different variables.

Table 4 provides the number of members who shares common characters i.e., the variables. This indicates a possibility to reduce the data using clusters and retrieve it in future using the imputation method without much loss in the common characteristics of original data. In what follows we discuss the statistical significance of the relationship which helps the reliability and validity of this approach. It may be noted that the clusters are determined at seven by observing the fact that most of the variables show a maximum decrease in these variables at this level of clustering. Since the data belongs to market research, the number of clusters that can be set is best when decided by the researcher himself based on his or her judgement of situations and requirements.

From Table 5 it is clear that there is a negative relationship with the number of clusters and standard deviation. To test whether there is a strong inverse relation a correlation has been applied to the data. The coefficient of correlation is strong with more than 0.6 when applied to all the experimental conditions. This is found to be significant at $p<0.05$. Therefore, when applied to more than two variables the correlation is found to hold well and is therefore an indication that this can be applied when more variables are present.

The correlation between standard deviation and number of clusters for two variables is noted as -0.930 and for the engine -0.870. Both these correlation coefficient shows highly negative nature of relationship. Further the correlation coefficient for three variable, for price and number of cluster is -0.968, engine -0.0798 and for horse power -0.873 as shown in Table 5. The negative relationship is confirmed with three variables as well. The correlation coefficient between number of clusters for four variables price, engine, horsepower and mpg are -0.973, -0.608, -0.833 and -0.865. The statistically significant negative relationship is confirmed as noticed in the Table 5.

Table 1: Pattern of standard deviation and number of clusters (two variable experiments)

	Cluster centroids		Standard deviation	
No. of clusters	Price	Engine	Price	Engine
2	36.75500	3.67500	12.2	0.9
3	33.07300	3.40300	9.3	0.6
4	33.53250	3.32750	7.5	0.6
5	37.78600	3.59400	6.2	0.6
6	35.28671	3.49996	5.5	0.6
7	35.28670	3.50000	5.7	0.4

Table 2: Pattern of standard deviation and number of clusters (three variable experiments)

	Cluster centroid			Standard deviation		
No. of clusters	Price	Engine	Horse power	Price	Engine	Horse power
2	28.05	3.110	188.8	10.1	0.7	36.7
3	34.85	3.500	213.4	9.3	0.7	31.3
4	31.21	3.253	198.4	7.5	0.6	24.7
5	36.86	3.514	216.4	6.1	0.7	28.3
6	36.56	3.485	216.3	5.1	0.6	27.0
7	41.39	4.053	246.6	5.2	0.5	19.3

Table 3: Pattern of Standard deviation and number of clusters (Four variable experiments)

	Cluster centroids				Standard deviation			
No. of clusters	Price	Engine	Horse power	Miles per gallon	Price	Engine	Horse power	Miles per gallon
2	27.39	3.055	185.1	23.85	10.6	0.7	37.4	3.304
3	34.41	3.483	211.5	22.73	9.7	0.7	31.3	2.555
4	29.62	3.120	190.0	24.57	7.8	0.6	28.6	2.860
5	29.09	3.186	190.2	23.87	6.8	0.6	28.6	2.304
6	34.61	3.442	208.1	23.20	5.5	0.7	30.9	2.338
7	39.72	4.016	239.5	22.27	5.6	0.5	23.2	2.164

Table 4: Cluster membership by variables

No. of clusters	Cluster member	Membership No. 2 variable	3 variable	4 variable
2	1	129	83	76
	2	26	72	77
3	1	63	72	70
	2	70	18	64
	3	22	65	19
4	1	51	54	19
	2	20	18	59
	3	60	44	59
	4	24	39	16
5	1	49	44	39
	2	13	39	54
	3	9	49	16
	4	60	7	25
	5	24	16	19
6	1	21	39	54
	2	11	44	16
	3	9	19	39
	4	57	32	13
	5	23	14	24
	6	34	7	7
7	1	8	39	54
	2	11	32	16
	3	1	44	13
	4	23	19	1
	5	34	14	6
	6	57	1	24
	7	21	6	39

Table 5: Correlation between number of clusters and standard deviation

Standard deviation	No. of clusters
Two variable	
SD of price	-0.930***(0.007)
SD of engine	-0.870**(0.024)
Three variable cluster	
SD of price	-0.968***(0.001)
SD of engine	-0.798(0.057)
SD of horse power	-0.873**(0.023)
Four variable cluster	
SD of price	-0.973***(0.001)
SD of engine	-0.608(0.200)
SD of horse power	-0.833**(0.039)
SD of miles per gallon	-0.865**(0.026)

*Pearson, values in the parenthesis show significance at 5%

Since the correlation does not show the cause and effect of the relationship, a linear curve estimation for each data set is performed. The independent variable is taken as the number of clusters and dependent variable as standard deviation. The linear equation tested is a+by where 'a' is a constant, 'b' is slope and 'y' is a variable in the linear relationship. The parameters estimated are given in figures.

Figure 1a explains the relationship between standard deviation of price and cluster numbers. The graph shows a linear relationship.

Figure 1b explains the relationship between standard deviation of engine and cluster numbers. The graph shows a linear relationship.

Figure 2a explains the relationship between standard deviation of horse power and cluster numbers. The graph shows a linear relationship.

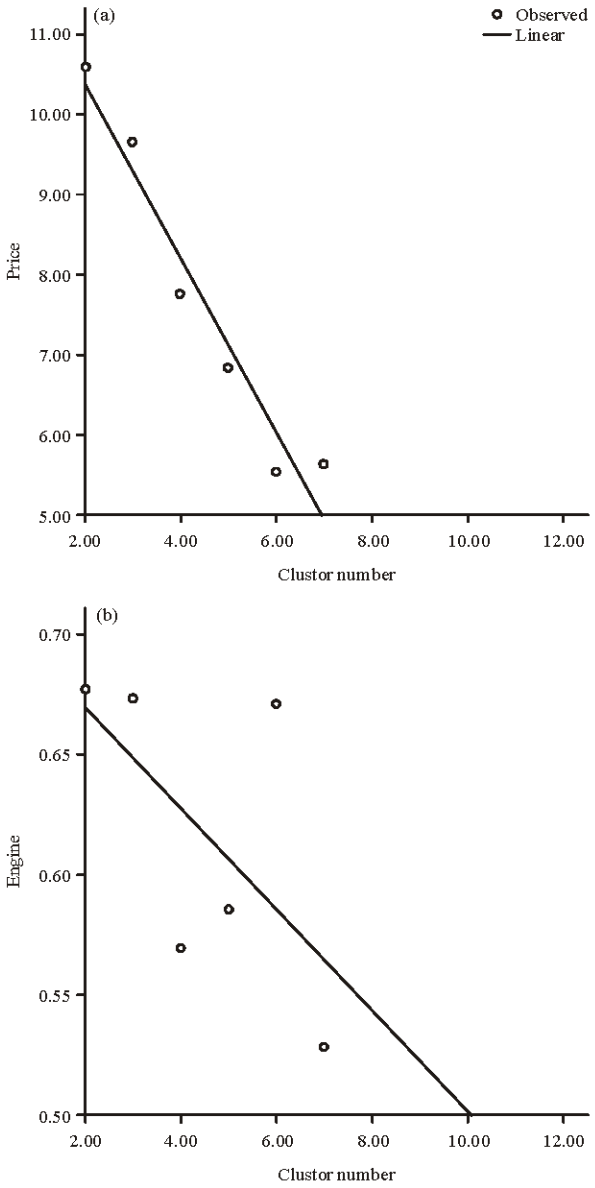

Fig. 1(a-b): Regression of standard deviation of (a) Price and (b) Engine on number of clusters

Figure 2b explains the relationship between standard deviation of miles per gallon and cluster numbers. The graph shows a linear relationship.

From the above estimates it is clear that the number of clusters is the cause for the increase or decrease of the standard deviations. All the slopes estimated show a negative relationship. The linear trend taken for variable price and engine is separately shown in the graph for two cluster experimental condition.

To test whether the pattern is linear or non linear it has been put to test using a regression analysis. Table 6 shows the parameters estimated with their corresponding significance.

All the models are tested using two variables i.e., number of clusters as an independent variable and standard deviation

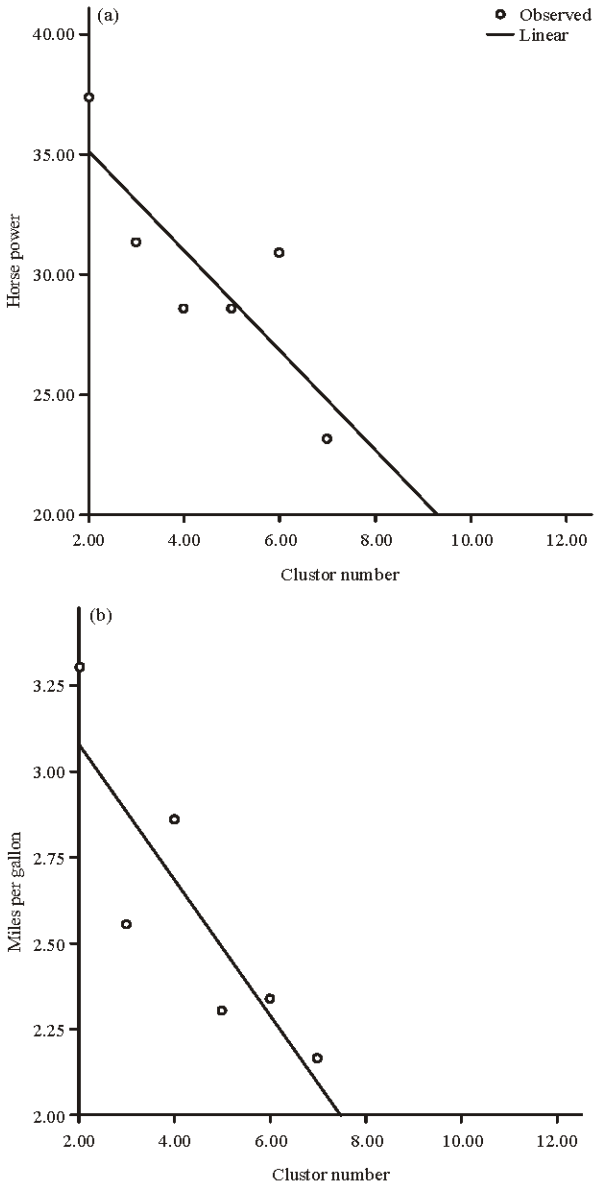

Fig. 2(a-b): Regression of standard deviation of (a) Horse power and (b) Miles per gallon on number of clusters

Table 6: Regression parameters

Variables	R^2	Sig.	B1	Constant
2 variable price	0.866	0.007	-1.296	13.561
2 variable engine	0.758	0.024	-0.067	0.914
3 variable price	0.938	0.001	-1.087	12.099
3 variable engine	0.636	0.057	-0.040	0.821
3 variable horse power	0.762	0.023	-2.756	40.283
4variable price	0.946	0.001	-1.087	12.560
4 variable engine	0.370	0.200	-0.021	0.712
4 variable horse power	0.694	0.039	-2.068	39.294
4 variable miles per gallon	0.748	0.026	-0.197	3.475

as dependant variable. It is found that all the models fit with a linear equation (a+b(y)) with high R^2. It can be seen from Table 6 that the slope of two variables with price and engine is -1.29 and -0.67. The negative relationship with high R square, points out the negative relationship between the variables keeping the number of clusters as an independent variable (a cause variable) and the standard deviation as a dependant variable (effect). Since all the slopes in Table 6 are negative it indicates a negative relationship and further confirmed the cause and effect relationship. Figure 1-2 confirm the fit and pattern graphically. Data simulation with more observations also showed similar results and provides sufficient accuracy. From the above analysis it is clear that the cluster centroids can be safely used to replace the original data. This outcome is very useful in compressing the data to save costly server space. However, the compressed data is useful in interpreting meaningfully when it is similar to the original data.

Test for differences between original data and centroid based data: The final stage of our experiment is to test whether there is any significant difference between original data and decompressed data. A paired sample t-test was used to determine whether there was statistically significant mean difference between the original data of price, engine, horse power and miles per gallon and data after replacing with cluster centroids in place of original data. Some outliers were detected which were a cause of concern and square root transformed data resulted in achieving normality and no outlier assumption for price, engine, horse power and miles per gallon which is evident from Shapiro_wilk p = -0.227, p = -0.261, p = -0.756, p = -0.686, respectively. After cluster in respect of price showed (27.444+13.33) as opposed to original data (27.44387+14.437), t(152) = -0.001, p = 0.999>0.05, d = 2.05, engine showed (3.0600+0.92545) as opposed to original data (3.059+1.0538), t(152) = -0.013 p = 0.990 > 0.05, d = 3.31, horse power showed (185.3963+52.19283) as opposed to original data (185.40+57.103), t(152) = 0.001, p = 0.999>0.05, d = 3.55, fuel efficiency showed (23.8276+3.78108) as opposed to o riginal data (23.83+4.293), t(152) = 0.011 p = 0.991>0.05, d = 6.30. A statistically no significant difference of -0.000510 in respect of price (95% -872251 to 87132), -0.00052 in respect of engine (95% -0.08080 to 0.07976), 0.00235 in respect of horse power (95% -3.69875 to 3.70346) and 0.00176 in respect of miles per gallon (95% -0.32346 to 0.32699).

From the test result it is evident that the cluster centroids can be used to replace the original data to compress the data. In future, cluster centroid based dictionary method can be used to decompress the data which very close to original. The method saves the server space and helps a lot in managing the data.

DISCUSSION

Zhang *et al.* (2006) contributed to and mentioned the use of clustering categorical data. However, this study focused on ratio scale data and thus it discusses the potential use of the market research data compression using clustering. The discussion done by Cilibrasi *et al.* (2004) is on the use of compression of music based on string compression which is

different from market research data that is discussed in this particular study. The previous studies on image compression and speech compression are discussing aspects and methods that are different from the use of cluster based compression on the market research data and ratio scales. The present study analyses the use of clustering in compressing the market research data and the fact that the more clusters results in lesser deviation among the cluster members. Table 1-3 are evidences for the same. It also tested and found that the similarity of data between original and cluster centroid based data have no statistically significant dissimilarity. Therefore the study indicates that the cluster centroid based compression can be used for decoding the original data without any significant loss. Moreover, the present study throws open the potential use of the technique in developing mobile applications based on dictionary based method using cluster centroids which can easily adapt to the mobile devices and helps in saving server space.

CONCLUSION

The study has unveiled one clear implication that more the cluster, the less will be the deviations from the original data in each cluster. The standard deviations examined for different variables found to be negatively correlated to the number of clusters. The test results proves that the cluster centroid can be used to replace the original data as it is very close to the original data showing no statistically significant difference. This indicates the possibility of compressing huge behavioural, categorical and ratio scale data sets without many deviations from the original by replacing with centroid. It identifies the common attributes of subset of groups in the large data set minimising the loss of original character. It is also found that the nature of relation between the number of clusters and standard deviation tends to be a linear one. Future researchers can find the existence of a non linear pattern with larger set of data and to explore an efficient method to decrypt the cluster centroids using dictionary method. It must be understood that the data set which cannot be meaningfully clustered cannot be used for compressing the data for the larger purpose of data compression. Even though this is a limitation it contributes largely to the likely usage of cluster centroid as a method for compression and points towards the future usage of centroid based dictionary method for encoding and decoding the data to save server space. This also indicates the possibilities of using the technique to access data conveniently using mobile devices.

REFERENCES

Cerra, D. and M. Datcu, 2012. A fast compression-based similarity measure with applications to content-based image retrieval. J. Visual Commun. Image Represent., 23: 293-302.

Cilibrasi, R., P. Vitanyi and R. de Wolf, 2004. Algorithmic clustering of music based on string compression. Comput. Music J., 28: 49-67.

Friedman, J.H. and J.J. Meulman, 2004. Clustering objects on subsets of attributes (with discussion). J. R. Stat. Soc.: Ser. B (Stat. Methodol.), 66: 815-849.

Lambert, D., 2003. What Use is Statistics for Massive Data? In: Crossing Boundaries: Statistical Essays in Honor of Jack Hall (Lecture Notes-Monograph Series, Volume 43), Kolassa, J.E. and D. Oakes (Eds.). Institute of Mathematical Statistics, Beachwood, OH., USA., ISBN-13: 9780940600584, pp: 217-228.

Perlmutter, S.M., P.C. Cosman, C.W. Tseng, R.A. Olshen, R.M. Gray, K.C.P. Li and C.J. Bergin, 1998. Medical image compression and vector quantization. Stat. Sci., 13: 30-53.

Sabater, C.R., P.C.A. Esteban, A.M. Iscar and A.L. Diez, 2004. Clustering to reduce regional heterogeneity: A spanish case-study. J. Popul. Res., 21: 73-93.

Slonim, N., G.S. Atwal, G. Tkacik and W. Bialek, 2005. Information-based clustering. Proc. Natl. Acad. Sci. USA., 102: 18297-18302.

Smith, D.E., 1995. Speech compression in attendant services: Analysis of a queueing system with delay-dependent service times. Oper. Res., 43: 166-176.

Zhang, P., X. Wang and P.X.K. Song, 2006. Clustering categorical data based on distance vectors. J. Am. Stat. Assoc., 101: 355-367.