Middle-East Journal of Scientific Research 23 (2): 214-218, 2015
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2015.23.02.22091

Dictionary Based Behavioural Data Compression: A Clustering Approach

[1]Biku Abraham, [2]Varghese Paul and [3]Nebu John Abraham

[1]Saintgits College of Engineering, Kottayam, Kerala, India
[2]Toc H Institute of Science and Technology, Kochi, Kerala, India
[3]Malayala Manorama Company Ltd., Kottayam, Kerala, India

Abstract: Behavioral data are increasingly being used these days in market research and planning. Generally these data are of large in size which necessitates an appropriate compression method to save server space. Dictionary based compression techniques are most suitable because it uses less computing resources. Generally these large set of data can be used effectively if it is accessible through mobile devices which has limited space. In this paper we explore a dictionary based compression using clustering method in respect of behavioral data. This paper uses dictionary based techniques as error of approximations are acceptable. The K means clustering techniques are used to generate the data in the dictionary. This work proposes two types of dictionary – primary and secondary, which are stored in the system instead of the original data set. For decoding, these dictionaries will be used so that the data set can be retrieved without losing any entry. This paper shows that this approach provides better compression ratio compared to conventional compression techniques with a very low root mean square error in comparison to the original data indicating minimal dissimilarity.

Key words: Compression · Clustering · Behavioral data · Lossy compression · Dictionary design

INTRODUCTION

Large volume of data is generated every day, everywhere. Data set analysis and data mining are becoming highly important because of its larger size. The data sets are becoming very large to easily transfer from one device to another. Everyday its gets appended and the size itself enlarges. Unfortunately, although the processing speed of computers and the capacity of onboard memory are both increasing at an incredible pace, the network bandwidth is not that much increasing.

Developments in computing devices such as tablets, mobile applications, navigation units and PDA's can be of great help to marketing professionals. These computing devices help sales persons to examine the behavioral position or attitude of a customer before approaching him. Search using these devices can identify customer behavior pattern approximately. Though this is lossy, this may perhaps be of great help to a marketing professional to interpret situations. This lossy information will be helpful for sales planning and prospecting. Portable with limited capacity, computing devices require an access to large data sets to take a decision for planning and probing with a marketing objective. Tablets, mobiles and personal digital assistant (PDA) do not have large storage space and is having limited data transportation bandwidth. In such circumstances larger datasets can be carried in these devices using compression techniques.

In behavioral data, lossy compression schemes are acceptable as marketing managers use those data for an approximation of attitudes of various target groups and customers. This makes dictionary based compression techniques suitable, as the error of approximation can be controlled and limited to a larger extent. This paper presents a dictionary based compression technique using K means clustering method. Dictionary entries are proposed to identify using K means clustering. The study shows that this proposed dictionary construction yields better compression ratios than the conventional compression algorithms like Win Rar or WinZip. Retrieved data tends to be almost original data itself.

A cluster is defined as a collection of data objects that are similar to one another and thus can be treated collectively as one group. Clustering is known as unsupervised learning as there is no pre defined classes.

Corresponding Author: Biku Abraham, Saintgits College of Engineering, Kottayam, Kerala, India.

It is a process of grouping similar objects together. The quality of a clustering method is measured by its ability to discover all the hidden patterns. Thus clustering is a prominent data mining task which has its applications in different fields such as marketing, city planning, insurance and land use etc.

Related Work: Shashi proposed the use of clustering algorithm to develop a dictionary based vector map compression for the data set. When the dictionary size is fixed, it shows lower error compared with the conventional algorithms [1]. From the review of literature it is noticed that among all the compression algorithms, dictionary algorithms perform well because it needs only less computing resources. For compressing large collection of related files, pair wise delta compression method was applied. Cluster based delta compression shows better results in pair wise compression [2]. Clustering approach reduces the computing cost in compressing files. Nowadays all companies perform behavioral analysis as a part of the customer segmentation for promotional campaigns. Clustering is used for behavior segmentation which is based on usage rate, price sensitivity, brand loyalty etc [3]. Companies need to keep the data for behavioral segmentation. With the advent of digitization, huge volume of data is available with the enterprises and behavioral data compression is very essential in enterprises [4]. Clustering techniques can be applied in the field of behavioral analysis. Literature survey explores the advent of data compression using clustering in different types of data like high dimensional discrete attribute data sets [5], images, related files etc. K means clustering technique are widely used to partition the data sets [6]. Usage of compression, clustering and its applications on behavioral data is an unexplored area and hence the main objective is to examine this area.

Methodolgy Adopted: In this study we performed a lossy compression technique based on clustering method using a dictionary. The data set we used for this work is obtained from an authentic marketing research survey conducted among Malayala Manorama weekly subscribers to find the market potential and sales conversion using a huge random sampling method all over the state of Kerala in India. The size of the data is huge and contains large data set. The data set comprises of eighteen variables. These variables are described in table 1 and the samples include the scaling values one, two, three and four. One represents strongly agree whereas four represents disagree.

Table 1: Description of Variables

Variable	Variable description
V1	I watch TV serials daily.
V2	I don't miss reading because of TV.
V3	I travel a lot.
V4	I have lots of leisure time.
V5	I am interested in discount schemes when I buy things from shops.
V6	I think a lot before a purchase due to limited budget.
V7	I am a regular reader of news paper.
V8	I am an instant adapter of new fashions.
V9	I try new recipes.
V10	I love old Malayalam songs than new ones.
V11	Advertisements influences me while buy products.
V12	I am a self focused person look for my utility.
V13	I like old fashions than new ones.
V14	Songs, cinema, mimics and dramas are the best interest of me.
V15	I am not interested in politics.
V16	I don't buy new technology products.
V17	I am a brand loyal customer.
V18	I think luck is more important than hard work.

Table 2: Scaling points

Scale value	Description
1	Strongly agree
2	Agree
3	Partially agree
4	Disagree

The scaling points are shown in table 2. The database contains the response of 41,715 subscribers. The analysis was done in IBM SPSS 20. The results provided approximations of original data leading to lower error while taking marketing decisions. To analyze the similarity of original data and decompressed data we have calculated root mean square error using MATLAB.

In this paper evaluation is performed on a real data set with a dictionary using K means clustering algorithm. The concept of coding the dictionary using K means clustering algorithm is used only to assess the compression rates and there is no scope beyond this objective. Data set which include behavioral data are discussed and therefore data sets other than behavioral data are beyond the scope of this paper.

Behavioral Data Compression Issues and Problem Definition: Behavioral data is a psychological approximation of a customer and understanding the clusters of main attitudes which are worth looking at for an effective marketing decision. It may cover attitudes, perception, probability of actions, strategic behavior etc.

Each dataset have primary data and behavioral data. Behavioral data is mostly assessed using a 9 point scale or with a 4 point scale. Primary data like customer name, brand name, identification number may be retained and to be preserved while lossy compression techniques may be used for the approximation of the behavioral data. Users of market research data require less stringent accuracy than users of other types of scientific data. This paper defines the problem of designing a dictionary using K means clustering. To achieve this objective, the main constraint is to keep the primary data such as customer id or name.

Dictionary Based Compression Usingclustering Technique (DBCC): A clustering algorithm is used to generate a dictionary in the dictionary based compression method. In this study we have used K means clustering algorithm for dictionary design. K means clustering generates as many clusters for a given data set as we desire in such a way that can be used for closeness with the original data. Thus the obtained K means cluster centroid become the dictionary entries in our proposed method. Each data could now be assigned to a particular cluster. Therefore this data would now be represented as a terms of reference to the cluster centroid in the dictionary. The primary data we obtained earlier will also be used to decode the data set using this dictionary. For decoding purposes this dictionary is sent along with the encoded data.

Algorithm for Compression:

Step 1: Train the data set so that replace the missing value with the series mean.

Step 2: Divide the data set into n clusters using K means algorithm; where n is defined as the optimal number of clusters depending upon the data set.

Step 3: Cluster membership and centroid are stored in a dictionary for each variable.

Step 4: Cluster membership and case number are stored in another dictionary. This dictionary is compressed using Win Rar or WinZip software.

Algorithm for Decompression

Step 1: Map the cluster membership with the case number.

Step 2: Store the cluster centroid value in the corresponding cases where the cluster membership matches.

Step 3: Repeat step2 till the end of the data set.

Step 4: Generate the data set with these recoded value.

RESULTS AND DISCUSSION

The data set in Microsoft Excel includes the attitude of 41,715 customers from market research survey conducted by Malayala Manorama weekly, the largest circulated weekly magazine in India. The data contains the responses from all the districts from Kerala state in India. Dictionary based compression method is applied on this data set. The K means clustering was done on this with a value of K set to twelve to build a dictionary of twelve main entries. The research shows that when the number of cluster increases the deviation from the original data to the compressed data will be minimum [7]. The number of cluster is increased to 12 on the basis of the fact that standard deviation decreases with number of clusters. But this may be limited as per the judicial decision of data manager. The original data set was then encoded using the dictionary thus obtained. The dictionary is used to compress the original data set. The final step is to evaluate the compression rates by computing the ratio of original data size and compressed data size. This has been compared with Win Rar to evaluate the effectiveness of this method.

Number of cases in each cluster is explained in the Table 3. There are a total of 41,715 cases considered. Out of these, 3942 cases are in cluster one, 2548 cases are in cluster two, 3678 cases are in cluster three and so on.

Table 3: Number of cases in each cluster

Cluster Number	Number of cases
1	3942
2	2548
3	3678
4	2808
5	3419
6	2602
7	2682
8	3998
9	4850
10	4278
11	3772
12	3138

Table 4: A sample dictionary which contains behavioral data marked in a 1 to 4 point scale

	Number of Cluster											
Variables	1	2	3	4	5	6	7	8	9	10	11	12
V1	4	3	4	2	3	3	3	3	3	3	1	3
V2	2	1	2	1	1	1	1	1	1	1	1	2
V3	3	3	3	2	3	3	2	3	3	2	3	3
V4	3	2	3	2	3	2	3	3	3	3	2	3
V5	4	3	3	3	3	2	2	4	3	4	3	3
V6	3	2	2	2	2	1	1	2	2	2	2	2
V7	2	2	2	1	2	1	1	1	1	1	1	4
V8	4	4	4	2	4	3	3	4	4	4	4	4
V9	3	3	3	2	3	2	2	3	3	2	3	3
V10	2	2	1	2	3	2	2	3	2	1	2	3
V11	4	4	4	3	4	4	3	4	4	4	3	4
V12	4	3	4	3	4	4	3	4	4	4	4	4
V13	3	2	1	3	3	2	2	3	2	2	2	2
V14	2	1	1	2	2	1	1	2	2	2	2	2
V15	3	4	1	3	4	2	2	1	2	3	2	2
V16	4	3	2	3	3	3	3	3	3	3	2	2
V17	4	3	3	3	3	2	2	3	3	2	2	3
V18	3	2	4	3	3	3	1	4	2	4	3	3

Table 5: Change in the data set size (Pre and post compression)

Compression Technique	Pre Compression Size (In KB)	Post Compression Size (In KB)
DBCC	6,049 KB	389 KB
Win Rar	6,049 KB	738 KB

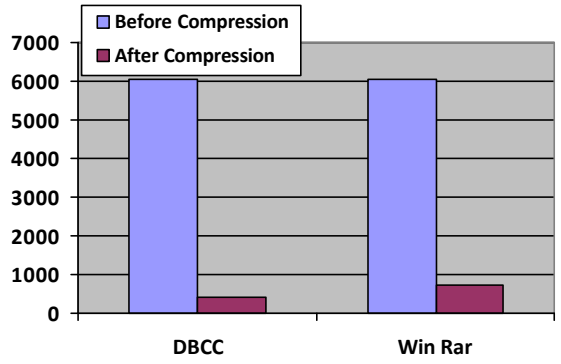

Fig. 1: Comparison between DBCC and Win Rar in terms of data set size.

A sample dictionary is shown below in Table 4. This dictionary contains eighteen variables and its cluster centroids. Each variable entry is divided into twelve clusters. The data set contains the attitude of 41,715 customers in four point scale. According to the behavioral pattern, these dataset are clustered into twelve divisions. V1 to V18 represents variable1 to variable18. Table 3 describes the number of cases in each cluster. For variable1 (V1), 3942 customers are in cluster one and their average opinion is marked as four. Average score of four represents "Disagreement" to the variable corresponding. Similarly for variable 2(V2), 2548 customers have similar opinion and they are in cluster one and the centroid is 2 and so on. This secondary dictionary along with the primary dictionary is used to decode the data set. The primary dictionary contains the cluster number and the case number. At the time of decoding, the cluster centroid will be allotted whenever the cluster membership matches. Cluster centroids described below represents attitudes.

The conventional method Win Rar is compared with the Dictionary based compression using clustering (DBCC) technique. The data set size of the file before compression and after compression is tabulated in the following table (Table 5). From the table it is clear that after compression, size of DBCC is only 52% of Win Rar. This shows a considerable improvement in the compression ratio.

Post compression size of DBCC is reduced by 52% of what Win Rar compression results provided. Therefore DBCC is significantly more effective than the conventional methods. The comparison of DBCC and Win Rar is represented graphically in the following figure 1. The y axis represents the size of the data set in kilo bytes and the x axis represents two types of compression techniques DBCC and Win Rar. The bar shows the size of data before compression and after compression.

Similarity Assessment of the Post Compressed Data in Comparison with Original Data: To verify the difference between the data before compression and after compression, the root mean square error of the data set is calculated. This data set shows an average of 0.02 for each variable in the data set. The root mean square error is 0 means both data are exactly same. Root mean square error of 0.02 means the difference between the values in the original and the compressed data set is very less. This result proved that dictionary based compression is very effective in behavioral data because market researchers are less stringent in the accuracy of their data.

CONCLUSION

The study explored an innovative method to compress behavioral data using dictionary based method. The K means clustering technique is used to generate the dictionaries. The primary (comprises of case number and cluster membership) and secondary dictionaries (comprises of variable number and cluster centroid) are stored instead of the original data. Decompression is done using these dictionaries. The main advantage of the DBCC method is that it reduces the post compression size to as low as 50 % compared to what Win Rar result provided. This study also found that the dissimilarity between the original data and the decompressed data is minimal as indicated by the low root mean square error. Since the market researchers are less stringent, even approximations are acceptable. Therefore this method of storing data is very useful to adapt it conveniently for sales persons as they could access information through various limited space mobile devices and other personal digital assistants. However, this study aims only on behavioral data and the work can be extended to other types of data like maps and images.

REFERENCES

1. Shashi Shekhar, Yan Huang, Judy Djugash and Changqing Zhou, 2002. Vector map compression: A clustering approach, ACM.
2. Ouyang, Z., N. Memon, T. Suel and D. Trendafilov, 2002. Cluster-based delta compression of a collection of files, in Web Information Systems Engineering, 2002. WISE 2002. Proceedings of the Third International Conference on, pp: 257-266.
3. Charu Nath, Rohit Kumar Akhairamka, Sanchit Bhatia and Varsha Ahuja, 2011. Behavioral analysis using data clustering, 1(1): 01-05.
4. Sanghamitra and Nilendra Chaudari, 2014. Scalable approach for analytics based customer segmentation, CSI Commun., 38(4): 16-19.
5. Koyuturk, M., A. Grama and N. Ramakrishnan, 2005. Compression, clustering and pattern discovery in very high-dimensional discrete-attribute data sets, Knowl. Data Eng. IEEE Trans., 17(4): 447-461.
6. Li, X., W.K. Cheung and J. Liu, 2010. Improving POMDP tractability via belief compression and clustering, Syst. Man Cybern, Part B Cybern. IEEE Trans., 40 (1): 125-136.
7. Abraham, B. and V. Paul, 2013. Compression of behavioral data using clustering technique, in Emerging Research Areas and 2013 International Conference on Microelectronics, Communications and Renewable Energy (AICERA/ICMiCR), 2013 Annual International Conference on, pp: 1-4.