BIOINFO Computational Optimization
Volume 1, Issue 1, 2011, pp-01-05
Available online at: http://www.bioinfo.in/contents.php?id=265

Behavioural Analysis Using Data Clustering

Charu Nath, Rohit Kumar Akhairamka, Sanchit Bhatia and Varsha Ahuja
[1]*Department of Information Technology, YCCE, Nagpur (MH)-441110, India*
e-mail: charunath@gmail.com, rohitakhairam@gmail.com, in.sanchit@gmail.com, cutevsa@yahoo.com

***Abstract*—In this paper, we discuss the feasibility of application of clustering techniques to perform behavioral analysis of people of organization. The organizations that more or less completely depend on its people for its productivity are the ones that constantly involve in analyzing the behavioral and attitudinal traits of its people. Two sound examples can be IT industry and educational institutes like schools and colleges. If employee of an IT industry and student of a college are in best of their behavioral attitudes in their domain, they produce the best results. Thus for such institutions, analysis of behavior has a pivotal role. At times such an analysis can be cumbersome and tiring. The task of analysis can be greatly simplified by the application of certain IT techniques that impose a classification and structure to data based on certain similar measures.**

To check the feasibility of application of clustering to behavioral analysis, we applied the clustering technique on final year students of Dept. of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur (MS), India. The students were made to fill a questionnaire that brought out their behavioral instincts. The instincts, represented in form of vectors, were fed to k-means clustering algorithm implemented in C. The results obtained contained clusters of students with similar behavioral traits. Now the behavioral traits of a student can be generalized to the cluster to which it belongs. Thus it cuts short the procedure of analyzing every student individually.

Keywords: K-means. Clustering, Behavioural Analysis

I. INTRODUCTION

Knowledge represents the intangible aspect of an organization. As process and product, knowledge of an organization is like a growing entity. Any organization can be found in a continuous loop of working over some knowledge, producing better versions of it and revising it for betterment of its subjects and society.

The paper primarily concerns of those type of organization where the most dominant part of asset base constitute human beings. The reason can be explained with the help of an example below.

Consider a production unit that assembles cars from its constituent parts. A few years down the line, there used to be one worker on one assembling unit that assembled one car at a time. For the production unit, the person working with the machine was equally an asset as the machine itself as the person was responsible for maintaining the machine and making it work. But as you see the current scenario one man maintains a machine that assembles 10 cars at a time. Obvious the membership of the man as asset to the organization has reduced and that of the machine has increased.

Now compare it to an IT industry. With advent of technology, number of people required by the industry has increased. Human beings still dominate the work in an IT industry, thus being it major asset.

Similarly in school and colleges the performance of every student affects the performance of the institute. Any thing like infrastructure or amount of facilities available cannot replace the factor of performance.

Thus these human intensive institutions always try out methods to improve and enhance the quality of their major asset human beings. Through different methods like Counseling, workshops, personality development classes. The human intensive organization tries to inculcate certain behavioural and attitudinal traits desirable in that organization.

But there are few questions that need to be answered.

How much time would it take to analyse the behavioural traits of every employee of an industry or every student of a college?

How do manager decide which people among hundreds require a personality development workshops and which requires Counseling?

Inspite of putting in efforts towards every individual employee or student, inspite of organizing workshops and training for people just because some key decision maker feels they may help.

We can use clustering technique to simplify the dataset. We are working upon and then do more logical and simplified decision making.

II. DATA CLUSTERING

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The main goal of clustering is to discover the groups or subgroups rather than to model them statistically.

In the applications of clustering, clustering techniques allow the division of groups or subgroups being done automatically, without any preconception about what kinds of groupings should be found in the pattern set being analysed. In data clustering, we basically perform five tasks viz. Pattern representation, selection of data proximity measure appropriate to the domain working in, clustering, data abstraction if

BIOINFO Computational Optimization
Volume 1, Issue 1, 2011, pp-01-05
Available online at: http://www.bioinfo.in/contents.php?id=265

required and result assessment if required. The output clustering can either be hard i.e. a partition of data into groups or can be fuzzy i.e. each pattern has a variable degree of membership in each output cluster.

Clustering techniques have been applied to a wide variety of research problems.

For example:

- In the field of medicine, clustering diseases, cures for diseases, or symptoms of diseases can lead to very useful taxonomies.
- In the field of psychiatry, the correct diagnosis of clusters of symptoms such as paranoia, schizophrenia, etc. is essential for successful therapy.
- In archaeology, researchers have attempted to establish taxonomies of stone tools, funeral objects, etc. by applying cluster analytic techniques.
- In general, whenever one needs to classify a "mountain" of information into manageable meaningful piles, cluster analysis is of great utility.

Clustering has also been used in a variety of other problems like

- Grouping of chemical structures.
- Search Result Grouping.
- Image Segmentation.
- Market Research.
- Social Network analysis.
- Data Mining.
- Grouping of shopping items etc.

A. Clustering Applied to Behavioural Analysis

If we are given a set of patterns or a set of feature vectors for some set of population then we would like to know if the data set has some relatively distinct subsets or not. In this context we can define cluster analysis as a classification technique for forming homogeneous groups within complex data sets. Typically, we do not know a priori the natural groupings or subtypes, and we wish to identify groups within a data set. We wish to form classifications, taxonomies, or typologies that represent different patterns in the data.

In cases where there are only two features, clusters can be found through visual inspection by looking for dense regions in a pattern set. But as in case of behavioural analysis distinct clusters may exist in a high-dimensional space. A general way to find the candidates for the cluster centres is to form n-dimensional histogram of data and find the peaks in the histogram. Another way is to randomly select the cluster centres from the given data.

B. Clustering Techniques

Clustering can be broadly classified into hierarchical clustering i.e. clustering process that produces a nested series of clusters by either splitting or merging them based on the similarities or dissimilarities and Partitional clustering algorithms identify the partition that optimizes a clustering criterion.

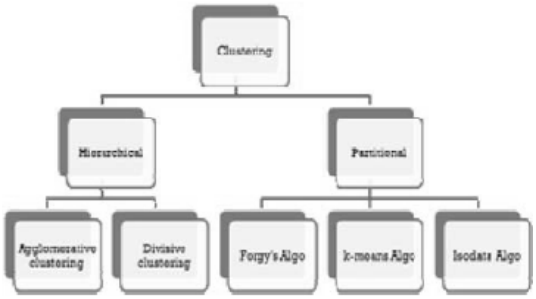

Fig. 1: Classification of Clustering Techniques.

1. Hierarchical clustering algorithm

Hierarchical clustering refers to the process that organizes data into large groups, which contain smaller groups, and so on in a tree structure. A hierarchical data clustering algorithm yields a multi-level dendrogram. It is called agglomerative if it is build from bottom up and it is called divisive if it is build from top down.

In Agglomerative clustering algorithm each pattern is in different cluster and clusters are successively merged till a stopping criterion is satisfied. In Divisive clustering algorithm all patterns are in one cluster and successively split till a stopping criterion is satisfied.

2. Partitional clustering algorithm (k-means algorithm)

A Partitional clustering algorithm obtains a single partition of the data instead of a clustering structure. The main goal of this algorithm is to create one set of clusters that partition data into similar groups and to group data that are close to each other. In many of these algorithms, the number of clusters to be constructed is specified in advance.

In k-means algorithm, besides data, input to the algorithm consists of k, the number of clusters to be constructed. It differs from the Forgy's algorithm in that the centroids of the clusters are recomputed as soon as a sample joins a cluster

General algorithm can be as follows:

- Choose k cluster centres to coincide with k randomly selected patterns inside the hypervolume containing the pattern set.
- Assign each pattern to the closest cluster centres.
- Recompute the cluster canters using the current cluster memberships.

If a convergence criterion is not met, go to step 2. Typical convergence criterions are: no (or minimal) re assignment of patterns to new cluster centres, or minimum decrease in squared error.

Fig. 2: Flowchart of k-Means Clustering

C. *Distance Measure*

The clustering method uses the dissimilarities (similarities) or distances between objects when forming the clusters. Similarities are a set of rules that serve as criteria for grouping or separating items. These distances (similarities) can be based on a single dimension or multiple dimensions, with each dimension representing a rule or condition for grouping objects. The most straightforward way of computing distances between objects in a multi-dimensional space is to compute Euclidean distances. If we had a two- or three-dimensional space this measure is the actual geometric distance between objects in the space (i.e., as if measured with a ruler). However, the Clustering algorithm does not "care" whether the distances that are "fed" to it are actual real distances, or some other derived measure of distance that is more meaningful to the researcher; and it is up to the researcher to select the right method for his/her specific application.

1. *Euclidean distance*

This is probably the most commonly chosen type of distance. It simply is the geometric distance in the multidimensional space. It is computed as:

Distance(x, y) = $\{\sum i \ (xi - yi)\ 2\}1/2$

Note: Euclidean (and squared Euclidean) distances are usually computed from raw data, and not from standardized data. This method has certain advantages (e.g., the distance between any two objects is not affected by the addition of new objects to the analysis, which may be outliers). However, the distances can be greatly affected by differences in scale among the dimensions from which the distances are computed and consequently, the results of cluster analyses may be very different.

3. *Squared euclidean distance*

We may want to square the standard Euclidean distance in order to place progressively greater weight on objects that are further apart. This distance is computed as:

Distance(x, y) = $\sum i \ (xi - yi)\ 2$

III. OUR WORK

A. *System Design*

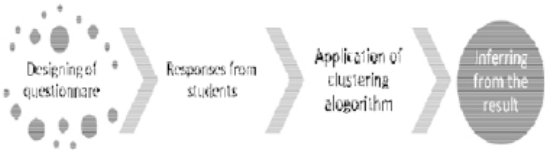

Fig. 3: Our Approach

Stage 1: Questioner was desidned by consulting different experts and interacting with students.
Stage 2: Responses collected from students
Stage 3: Response of the students converted into feature vector and supplied to the clustering routine.
Stage 4: Inferring the result obtained.

B. *Questionnaire Design*

The first step towards effective implementation was to design the questionnaire in such a way that clustering could be performed efficiently. As clustering of final year students was being done, their behaviour was observed, and with the help of a behavioural psychologist and performing field work like interacting with students and conducting interviews, a questionnaire was developed.

Some of the features which we came across after consulting various psychologists, counsellors and our own experience as students were:

Physical attribute, personality, family background, attitude, interest, curiosity, openness, practicality, independence, self-confidence, non judgemental, colour, food habits, etc.

BIOINFO Computational Optimization
Volume 1, Issue 1, 2011, pp-01-05
Available online at: http://www.bioinfo.in/contents.php?id=265

After considering the above features, a questionnaire was developed which was divided into three sections.

The first part of the questionnaire consisted of questions describing a student's personality and their approach to life. Questions like preference of colour, attitude towards life, their best quality according to them, dressing style etc. were some of the factors which helped in analysing their behaviour. Each question described a certain aspect of one's personality. e.g. For colour preference if one chose black/red it signified that he/she is bold , confident and open minded. Pink/Blue signified that he/she is sophisticated; white/beige represented the playing safe attitude and only black represents pessimism. A situation was also given and their reaction to that particular situation was gauged. E.g. if you were walking alone in the forest which object would you stop to pick up? If the answer was pen knife then the student is always trying to protect himself i.e. being defensive. If the answer was cell phone then the student has a practical approach towards life and if the answer was peculiar shaped grey rock then the student has a creative mind.

The second part of the questionnaire, emphases on relationships and a student's attitude towards relationships have been laid upon. Handling friends and other relations is an important factor in behavioural analysis. The questions covered in this part consisted of features related to understanding an individual, the various qualities seen in a person while making him/her your friend, confrontation with the opposite sex etc.

The first option implies that the individual is shy. The second option implies that he/she is street smart and can handle any situation. The third option indicates that the individual speaks his/her mind and the fourth option indicates that the individual always thinks before speaking and will talk very less.

The third part of the questionnaire consisted of questions related to academics .In these questions, a student's outlook towards studies and teachers were analysed. Questions on internal examinations, practical, self study, dependence on teachers were asked to check their level of independence, curiosity and responsibility. Questions related to height and aggregate marks were also asked as these features helped in intensive clustering.

Students having common behavioural traits were clustered together. Common attributes and attitude could define a group but not friends.

C. Implementation

Program to implement K-Means algorithm was written in C. Then it was executed over two test patterns in 2-Dimentional Space. Favourable results eliminated the notion of Improper working f the algorithm.

Then we applied the same algorithm over the behavioural data obtained from the final Year students of department of information technology of our college. The responses of the students to the designed questionnaire were manually converted into feature vectors and fed to the program.

No. of Responses taken: 78

No. of Centre chosen: 10,12,14,16

The program created a file that stored the pattern vectors. The program read the data from the file and performed clustering. The clustering was done with taking 10, 12, 14 and 16 cluster centres respectively and inferences were drawn from the results obtained.

On successive implementations and with varying the number of cluster centers, it was observed that there exist certain patterns that appeared again and again in the results. Some of these patterns are:

- 1 31 67
- 71 72 73 80
- 12 20 49 78 79
- 50 51 58
- 8 63 66 69

These numbers represent the roll numbers of students with similar behavioral traits.

After seeing the variations in the results we discovered some problem with the questionnaire. One of the problems was the uneven numbers of options available in different questions. To counter this, we made some modification by normalizing the input feature score in the range of 0-1.

IV. CONCLUSION

Most favourable results were not obtained as students who once came together in one execution did not necessarily showed up in one cluster in another execution. But the traces of patterns found (as mentioned in observations) were a positive sign indicating the acceptable feasibility of clustering techniques to the field of behavioural analysis. Further the choice of clustering algorithm can be commented on. The data set we considered in our study would have worked well with fuzzy c-means clustering as it takes care of variable membership of the patterns into clusters.

It can, therefore be acceptably assumed that the clustering techniques can favorably be applied to the field of behavioral analysis and therefore can prove to be of great use to human-intensive institutes like IT industries and educational institutes.

ACKNOWLEDGMENT

We are extremely thankful to Prof. K.K. Bhoyar (HOD-IT, YCCE), Mrs Vani Rawat (Behaviour psychologist), Mr. T. Ram Mohan (Director, P.T.Education) and Mrs.

Poonam Maneria (Career counsellor) who provided their input at various junctures of our project and worked relentlessly with us.

We would also like to thank the students of 7th semester I.T. for giving valuable input to our project by filling up the questionnaire. We would like to extend our thanks to the teaching and non teaching staff of our department and also to the students of our college who directly or indirectly influenced and helped us in our project.

REFERENCES

[1] A.K. Jain, M.N. Murty and P.J. Flynn, "Data Clustering: A Review", ACM Computing Surveys, vol. 31, No. 3, 264-323. 1999.

[2] F.H. Borgen and D.C. Barnett, "Applying Clustering Analysis in Counseling Psychology Research", Journal of Counseling Psychology, vol. 34, No. 4, 456-468. 1987.

[3] R. Dass, "Using Association Rule Mining for Behavioural Analysis of School students: A Case from India", IEEE 42ND Hawaii International Conference on System Sciences, 2009.

[4] R.O. Duda, P.E. Hart and D.G. Stork, Pattern Classification, 2nd edition, New York, USA: John Wiley and Sons Inc., 2001.

[5] E. Gose, R. Johnsonbaugh and S. Jost, Pattern Recognition and Image Analysis: Pearson Education, 1996.

[6] J. Canfield, J. Switzer, The Success Principles: Harper Collins Publishers Limited, 2005.

[7] (2009) The Wikipedia website. [Online]. Available: http://www.wikipedia.org.

[8] (2009) The ACM portal. [Online]. Available: http://portal.acm.org.