American Journal of Software Engineering and Applications
2016; 5(3-1): 20-24
http://www.sciencepublishinggroup.com/j/ajsea
doi: 10.11648/j.ajsea.s.2016050301.15
ISSN: 2327-2473 (Print); ISSN: 2327-249X (Online)

Analyzing Personality Behavior at Work Environment Using Data Mining Techniques

Sepideh Ahmadi Maldeh*, Fateme Safara

Faculty of Computer Engineering, Islamic Azad University, Islamshahr Branch, Terhan, Iran

Email address:
Sepide_am70@yahoo.com (S. A. Maldeh), safara@iiau.ac.ir (F. Safara)
*Corresponding author

To cite this article:
Sepideh Ahmadi Maldeh, Fateme Safara. Analyzing Personality Behavior at Work Environment Using Data Mining Techniques. *American Journal of Software Engineering and Applications*. Special Issue: Advances in Computer Science and Information Technology in Developing Countries. Vol. 5, No. (3-1), 2016, pp. 20-24. doi: 10.11648/j.ajsea.s.2016050301.15

Received: September 12, 2016; **Accepted**: September 17, 2016; **Published**: October 20, 2016

Abstract: Character is the influencing factor of human behavior. This research aims to analyze the relationship between different types of characters. The statistical society sample for this study is the employees of the Iran Mahd Parta Pajhohan technical complex. Two hundred employees have been divided into four clusters including: Type D (Dominant), Type I (Influential), Type S (Steady) and Type C (Conscientious). The analysis of the data has taken place at two levels, which are known as descriptive and inferential statistics. K means algorithm has been used to cluster employees, and as a result, most of the employees are DC personality types. The results help in improving the operation of the organizations as well as leading a healthy relationship between employees.

Keywords: Character, Employee, Data Mining, Clustering, Kmeans

1. Introduction

Research on human activity is complicated. There is a model that controls the behavior using statistical methods and machine learning. Based on the behavior, we take into consideration a part of individual characteristics known as character. The type of an individual's character is one's behavioral pattern or one's character is the influencing factor of his relation with others [1-2]. One's character reflects his way of thinking and behavior. Human beings express their opinions by their beliefs, feelings and behavior [3]. Scientists have done research on this topic in different angles. On the visual basis, special characteristic features like capabilities, values and one's credibility are the influencing factors. Today receiving public opinions about social events, marketing activities are the interests of the society. The Opinion Mining is to discover and restore the information from the world network. They have used some learning methods that are without supervision, a dictionary of feelings has been made and used to decide the degree of positivity and negativity of word or sentence, and they have made a labeling algorithm. In addition, some studies have used the topical information of dependent relations to receive the product's information [4]. The analysis of feelings is used more in English languages but the processing of natural languages has been used in Indian languages, too [15]. The data mining concept is on the basis of social networks. In this research, we have tried this analysis on humans and by recognizing one's characteristic features we can have a positive effect on relationships with others and by using mining data algorithms we can predict their future behaviors, which in some studies have been used to predict the future behaviors of some students [5]. In this method, there is no good or bad behavior pattern. Research has shown that the most successful people are the ones who recognize their own tendencies and types of behavior and then by recognizing other's tendencies by taking into consideration the situation they are in company can use these strategies for having a dynamic relationship. This information is useful only when it is reviewed and debated. It is used as a process to increase the individual influence. DMT is a branch of artificial intelligence and has the most important initiative in the computer systems since 1960 [6]. The main methods in DMT have two categories: descriptive and predictive. Not much research has been done on data mining of social behavior from the year 2000 to 2011.

Qualitative questionnaires and statistical methods are the basis of researches in DMT that are mostly used in the concept of behavior recognition [7]. In this article, the following topics are studied.

In this article, the following topics are studied. Firstly, a revision on primary definitions and personality studies. Also the steps of data preprocessing and clustering algorithm and related factors. Finally, the suggested method has been concluded.

2. Primary Definitions

2.1. History of Mining and Character Studies

Character has been studied on different viewpoints. Some of them have been studied on the aspect of traits, some others on the aspect of psychology and some groups on the aspect of biology, some of them on the aspect of social and humanity. Also, the character type A&B was introduced for organizational management [3]. Today the psychology type DISC is a popular method in behaviorism that we will explain briefly.

DISC has four letters: D: dominance I: influence S: stability C: commitment. These four letters reveal the four types of personalities. The common features of DI: active, fast, firm, bold, extroverted. The common features of SI: pleasing, tactful, acceptant, compatible. The common features of CS: thoughtful, moderate, calm, caring, introverted. The common features of DC: interrogative, logical, skeptic, challenger, committed [8].

2.2. Data Mining

Data Mining is a process of discovering knowledge and is a branch of artificial intelligence. It is evaluating big volume of data inside the database. Methods of the data mining have two categories: the predictive methods that include regression, classification, dependent rules and the descriptive methods include dependent rules, discovery of educative patterns, classification, and analysis of the remote areas [9].

2.2.1. Missing Data

The primary data that we have for the data mining may not have certain features, for example: there is nothing mentioned about sex and age in the data. These are called missing data. Missing data may occur due to the following reasons:

1. The equipment may have problems.
2. It may be incompatible with the other data and not be mentioned.
3. It may not have been of any importance when the data was given.
4. Changes were not recorded in the history [10].

In this research, there were a lot of missing data. The fields where the sex and age were not filled were deleted and the fields of the states or places of living that were mostly Karaj area were not taken into consideration as they didn't make much difference in defining one's character. The fields that were predictive were written manually.

2.2.2. Confusing Data

Confusion or random error or conflicts have been calculated randomly and these data have an effect on calculations and they result in an incorrect pattern. The factors may be incorrect due to the following reasons:

1. The problems occurring during data input.
2. Technology limitations.
3. Faulty equipment of gathering data [10].

2.2.3. Dimension Reduction

Dimensionality of the data has an important effect on the results. PCA is one of the best linear techniques to reduce dimensions. In this method, there are fewer missing data. PCA is of course not limited to decreasing dimensions only. The amount of principle components that is produced should be similar to the main data. In this method, new orientations have been defined for the data, and the data are defined by these new patterns. Table 1 PC output is the Rapid miner software. The amount of PC that has the most similarities to the main data is selected [11]. PCA10 with 37% similarity and PCA8 with 31% similarity and PCA6 with 25% similarity have been studied in this research. In concern with the available data that include 112 variable data and 142 data, the PCA8 would be appropriate. Be careful that a lot of similarity to the main data is not needed. Similarities should be acceptable.

Table 1. *The matrices of PCA that show in table.*

Component	Standard Deviation	Proportion Variance	Cumulative Variance
PCA_1	2.041	0.058	0.058
PCA_2	1.847	0.047	0.105
PCA_3	1.769	0.043	0.149
PCA_4	1.672	0.039	0.188
PCA_5	1.488	0.035	0.223
PCA_6	1.581	0.035	0.257
PCA_7	1.514	0.032	0.289
PCA_8	1.443	0.029	0.318

In fact a paper moves among the data and the image of the data and the algorithm becomes clear [12].

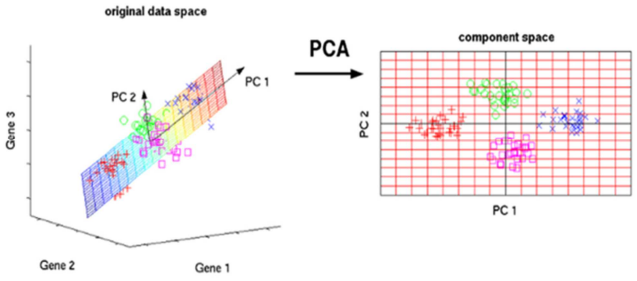

Figure 1. *Shows a more accurate image of the PCA [13].*

2.3. Clustering

The clustering algorithm divides the properties differently. These different groups are called clusters. Clustering is one

of the techniques of the learning with un-supervision mode. There is a cost criterion, which can be used as Euclidean, or Bergman Divergence [10]. In this article, we use the Bergman Divergence criteria and the K means algorithm.

2.3.1. The K Means Algorithm

The K-means is a clustering method that works as follows [10]:

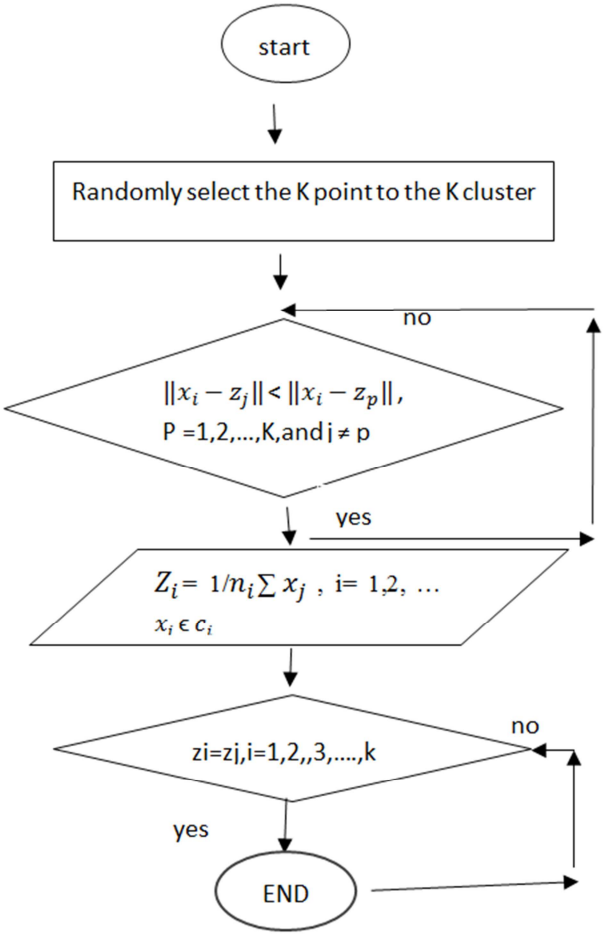

Figure 2. *The K-means clustering flowchart.*

2.3.2. Optimized Cluster

Evaluating the cluster is the most difficult process of the clustering analysis. If all the variations are independent, no clusters are formed. In that case, all the data form one cluster. In a situation between independence and full dependence, we will not know how many clusters are formed. Calculating the amount of K is the most important phase of clustering [10]. To achieve an optimized clustering, we use the SSE (Formula 1). SSE calculates the space within a cluster i.e. it calculates the distance between the sample and the centre of its cluster. The SSE method is as follows:

$$SSE = \sum_{i=1}^{K} \sum_{x \epsilon C_i} \text{dist } 2(m_i, x) \quad (1)$$

X is a point from Ci and Mi cluster of the centre. In SSE if a sample is kept within a cluster then SSE=0. A good clustering with a lower K has a lower SSE. There is another method called Davis Bolding which works like the SSE. In addition to calculating the distances within the clusters it calculates the distance between different clusters. The distance starts to become lower till it increases in one point and then continues to decline (Formula 2). As a result, the point where it increases, we get the optimized cluster [12].

S =1– a/b if a<b, (or s = b/a -1 if a≥ b, not the usual case) (2)

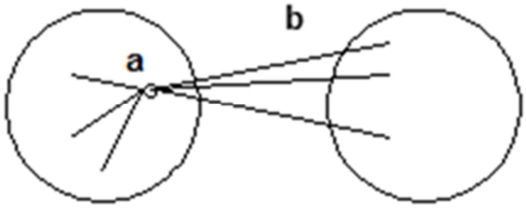

Figure 3. *a = the average of one sample and the others that are within a cluster and b = the average distance between this and the other samples [12].*

2.4. The Used Technologies

SQL is supported by RDBMS and is developed through Oracle [13]. The data of this article is saved in the SQL. One the data mining open text equipment is Rapidminer. Data mining organizes the learning machine, which includes loading data, transferring data, decreasing dimensions or dimension reduction, making the model, and evaluating it. This program is written with java script. The WEKA learning machine is used in the learning of patterns and evaluating the variations. The rapid miner program is used to produce patterns from K-means algorithm [14].

3. Methodology

In the behavior analysis, the data are collected from the website of Mahd parda Pajhohan technical institute. In this part, the implementation process is stated as follows:

Firstly, the collected data are explored and then the most important phase, which is the preparation of the data is done. In the clustering, the number of variations should be less so that the decreasing dimension would have happened that are explained later. Finally, implementation of the K means algorithm and a brief explanation of the result are given. This data collection includes 200 records, which are taken in order to study the behavior of the employees Mahdparda Pajhouhan technical institute and it includes 112 variations, which are: first name, last name, education, age, prospective, thoughtful, wise etc. This experiment includes 28 questions that are divided into 4 groups and each participant gives two answers to each question. One answer is given according to the most behavioral pattern and one as the least. Therefore, a score is given to each choice. The number zero shows not chosen ones and number one shows the least behavioral pattern and number two shows the most behavioral pattern. In the education field the textual data are transformed to discrete data i.e. number 1 is given to primary education, number 2 diploma, number 3 associated degree, number 4 BA, number 5 MA and number 6 given to PhD.

3.1. Preparation of Data

Mostly the collection of raw data includes special errors: incomplete data, confusing data, incompatible data, duplicated data, Null data, they are extracted from the database by using the Query. The parts where all the fields were null were deleted. Some of the predictive data were manually filled using the data abundance.

Clustering is usually used to analyze the data described with its features. Number of features has a significant role in the clustering performance. PCA is one of the processes of decreasing dimension. Each PCA is the outcome of the records. So PCA equals to 5 in the beginning. The Rapid miner software that produces PCA has 2 outputs, one is exa and the other is pre. The pre output provides a table with which we can find out the number of PCA. In figure 2 the estimation table shows the number of PCA. According to table 1 PCA8 is 30% similar to the main data and PCA6 is 25% similar, so according to the number of data and variables, PCA6 is more suitable.

3.2. K-means Algorithm

Clustering of the data is the most important part. The fewer the number of clusters, the better the data analysis. However, it is better that the number of clusters are optimized. This could be done in best way by using the SSE or the Davis bolding [12]. In this article Davis bolding is used. As a result, the optimized cluster that is PCA8 is achieved as follows: firstly, the number of cluster is kept as K=2 and then the Davis bolding equals to 2.57. Then we keep the number of clusters as 3 and the Davis bolding output is 2.14. This process continues until at one point the amount of Davis bolding increases. On the point where k is equal to 7 the Davis bolding equals to 1.52 and at the point where k equals to 8 Davis bolding equals to 1.61. Therefore, according to Davis bolding when the number of cluster equals to 7 the process of decreasing continues until the number of cluster becomes 8 and there is a sudden increase. Therefore, with the PCA8 the number 7 is optimal. But for the data analysis lower number of cluster is better therefore the number of PCA is decreased to 6 with a 26% similarity to the main data and according to Davis bolding the number of optimized cluster started at k=2 and 5 clusters and the number of optimized cluster is found.

Table 2. *The estimating process of finding the optimal cluster.*

# of clusters	Davies Bouldin Index with PCA8	Davies Bouldin Index with PCA6
3	2.14	1.99
4	1.72	1.61
5	1.66	1.33
6	1.55	1.36
7	1.52	1.31
8	1.61	1.29
9	1.44	1.26

Table 3. *Cluster Model with PC8, the optimal number of clusters is 7.*

Number of Cluster	Members
$Cluster_0$	14
$Cluster_1$	33
$Cluster_2$	20
$Cluster_3$	18
$Cluster_4$	23
$Cluster_5$	29
$Cluster_6$	5

Table 4. *Cluster Model with PC6, the optimal number of clusters is 5.*

Number of Cluster	Members
$Cluster_0$	40
$Cluster_1$	25
$Cluster_2$	5
$Cluster_3$	47
$Cluster_4$	25

4. Implementation and Results

According to the achieved conclusion from the model, we can say every variable is an average of each cluster. If it exceeds 1 it has a stronger behavioral pattern of a particular characteristic. According to table 3 the Clustering has 5 clusters. The members of group 2 did not answer any questions and the number zero is given. Therefore, their average equals to zero as a result this cluster is noticed. The characteristics of cluster zero are as follows: cluster zero with an average of more than 1: casual 1.25, strong 1.42, brave 1.4, determined 1.2, persistent 1.4, astute 1.15, competent 1.6, risk-taking 1.4. The average of characteristics that are lower than 1: hasty 0.8, thoughtful 0.8, obedient 0.95, intimate 0.95, grateful 0.85, shy 1, wise, 0.87, stimulating 0.8, open to other's feelings 0.97. These people have the type D characteristics that are powerful, strong, frank and bold, penetrating, competent, challenger, risk taking and are not obedient and lose their temper quickly. For being more effective these people should be given difficult tasks, they need no operate arts and methods practically using their experiences. They need to receive shocks at times. If they achieve results, they have to express it; they should realize they are needed by others. The characteristics of cluster 1 members with an average of more than 1 are: quiet 1.3, associable 1.25, obedient 1.25, spontaneous 1.16, moderate 1.5, historian 1.14, methodical 1.44, frank 1.25, patient 1.7, strong 1.4, healthy 1.3, brave 1.7, peace lover 1.12, calm 1.12, I spread peace 1.32.

Characteristics of members of cluster 1 with an average of less than 1 are: receptive 1, demanding and strict 1, risk taking 0.97, moody 1. These people have the S type of personality properties. They are calm, happy, popular, generous, romantic, and logical and of course slow. Resistant to changes and show no interests in risking and challenging. To be effective they need to keep up with the

situation before any changes. Their values should be appreciated. They must how much individual efforts can have an influence on group efforts. They should be loyal. Their creativity should be appreciated. The characteristics of members of cluster 3 with an average of more than 1 are as follows: express feelings 1.25, neutral 1.25, moody 1.7, talkative 1.25, perfectionist 1.31, frank 1.7, clear 1.7, contestant 1.6, eloquent 1.4. The people that are in cluster 3 have the DC type of personality. As you can see, there are 47 members in cluster 3. Therefore, most of the employees are DC type. The characteristics of members of cluster 4 with an average of more than 1 are norm acceptant (I obey rules and standards) 1.4, assessor (I evaluate everything) 1.6, cheerful 1.44, stimulating 1.44, disciplined 1.3, circumspect 1.44, astute 1.23, analyzer 1.34. The characteristics of members of cluster 4 with an average of more than 1 are moody 1.0, smart 1.0. These people have the personality type C (committed). They are disciplined, accurate, perfectionist, truthful, analyzer and have high standards. Actually, these people are critics of themselves and others. To be effective they need to do the planning carefully, to understand the targets and the explanations of the given missions. They should have plans to estimate their activities. They should receive the special feedbacks of their activities and obey the individual values as well as their achievements. They should have more patience in disputes and problems.

5. Discussion and Conclusion

In this article, we clustered the behaviors of people using data mining algorithm. Employees are divided into 4 groups according to their behavioral patterns that include personality type D, who are domineering, type I who are influencing, type S, who are stable and type C, who are committed. According to the achieved model, most of the employees have a DC type of personality. This clustering provides an appropriate solution to improve the activities of organizations in order to attract and protect the useful troops. This helps the organizations to improve their relationships with the employees as well as the managers within an organization by having knowledge of each other's behavior. In addition, by recognizing customer's personality types and analyzing selling methods, a more suitable atmosphere for business and relations with customers could be established and a more suitable work atmosphere could be created for the employees.

References

[1] Burche, A, Chandak, M. B, Zadgaonkar, A, Opinion Mining And Analysis: A Survey, International Journal on Natrual Languega Computing (IJNLC). Vol. 2, No. 3, June 2013.

[2] Liao, S-H, Chu, P. H, Hsiao, P. H, Dat mining techniques and application – A decade review from 2000 to 2011, Journal hompage: www.elsevier.com/lacate/eswa.

[3] Tehran publishing rasa Albvty. Mostafa ،Karl El Cooper and Raja calimo, Management of Socio_psychological Factors at work, Tehran publishing rasa, 2005.

[4] Bruin, j. s, Cocx, T. K, Kosters, W. A, Laros, J, Kok, J. N, Data mining approaches to criminal career analysis in Proceeding of the sixth international conference on Data Mining (ICDM 06), pp, 171-177, 2006.

[5] Fodor, I. K, A survey of dimension reduction techniques, technical report, Lawrence National Laboratory, June 2002.

[6] Cambria, E, Schuller, B, Xia, Y, Havasi, C, New Avenues in opinion Mining and Sentiment Analysis, Published by the IEEE Computer Society. 2013.

[7] Adhatrao, K, Gaykar, A, Dhawan, A, Jha, R, Honrao, V, Predicting Students Performance Using ID3 And C4.5 Classification Algorithms, International Journal of International Journal Knowledge Management Process (IJDKP) Vol. 3, No. 5, September 2013.

[8] http:// persiansun.persianblog.ir/post/2015

[9] Bruin, j. s, Cocx, T. K, Kosters, W. A, Laros, J, Kok, J. N, Data mining approaches to criminal career analysis in Proceeding of the sixth international conference on Data Mining (ICDM 06), pp, 171-177, 2006.

[10] Khanifar, Hossein, Moqimi, Mohammad, Fatehi, Narges Sadat. Data Mining and Knowledge Discovery, University of Science and Industry, 2009.

[11] Fodor, I. K, A survey of dimension reduction techniques, technical report, Lawrence National Laboratory, June 2002.

[12] Tan. P. N, Steinbach. M, Kumar. V, Introduction to Data Mining, Addison-Wesley, 2005.

[13] MySql – the world most popular open source database, http://www.mysql.com/

[14] Rapid Miner, http:// rapid.com/content/view/181/190/

[15] sharma. R, Nigam. S, Jain. R, Opinion Mining In Hindi Language: A Survey, International Journal in Foundations of Computer Science & Technology (IJFCST), Vol. 4, No. 2, March 2014.