Automated Personality Classification

Abdul Wahid[1][ID: 193-15-2992], Ata-E-Elahi[1][ID: 193-15-2946] and
Samanta Sajjad Shraboni[1][ID: 193-15-3023]
[1] Daffodil International University, Dhaka, Bangladesh

Abstract. Personality, a characteristic way of thinking, feeling, and behaving. Personality embraces moods, attitudes, and opinions and is most clearly expressed in interactions with other people. Personality can be defined as a necessary element of a person's behavior. The way people interact with other people determines their personality. This report covers the topic of Automated Personality Classification – a system that analyses the personality of a user based on certain features using Data Mining Algorithms. In this report, we propose a system which analyses the personality of an applicant. This system will be helpful for organizations as well as other agencies who would be recruiting applicants based on their personality rather than their technical knowledge. The personality prediction results are based on Big Five Personality traits and the classification is done using Naïve Bayes Algorithm and Support Vector Machine.

Keywords: Big Five Personality, Naïve Bayes Algorithm, Support Vector Machine, Automated Personality Classification, Data Mining

1 Introduction

Personality is one feature that determines how people interact with the outside world. Personality can be defined as a necessary element of a person's behavior. Automatic personality classification uses data mining algorithms to analyze personality based on certain characteristics. The personality prediction results are based on Big Five Personality traits. These 5 traits are Extroversion, Neuroticism, Agreeable, Conscientious and Open personality. The project uses learning algorithms and data mining concepts to mine user characteristics data and learn from the patterns. We propose a system that analyses the personality of a person. This system will be helpful for organizations as well as other agencies that would be recruiting workers based on their personalities. This method is also beneficial to social networks moreover as various ad selling online networks classify user personality and sell more relevant ads.

Predicting personality by analyzing the behavior of the person is an old technique. This manual method of personality prediction required a lot of time and resources. Analyzing personality based on one's nature was a tedious task and a lot of human effort would be required to do such analysis. This traditional method of predicting personality would require a lot of time and was very limited in scale. Also, this manual analysis did not give accurate results while analyzing the personality of a user from their nature and behavior. Since analysis was done manually, it affected the accuracy of the results as humans are prone to biases and prejudices.

Data mining techniques are therefore used to study and analyses data and then identify any hidden patterns or information from a large data set. These techniques are used to mine user characteristics and then train the model accordingly to predict the personality of other users in the future. Using these techniques, we can analyze the personality of an applicant applying for a job in an organization which gives priority to one's behavior and personality rather than technical knowledge. Also, the applicant gets to know what all personality traits are in him/her and what all traits are missing. Thus, we can then guide him/her to develop those traits or to strengthen the other traits accordingly.

The main objective of this paper is to overview the data mining algorithms which are used to predict the personality of the user. In this paper, we focus on an online test which would be given by the applicant and then his/her personality would be predicted accordingly based on the Big Five Personality traits. In this way, we can filter out candidates applying for a specific position in the organization. Thus, it would save the resources of the organization and they would then interview only those candidates which would be most suitable for the job.

2 Research Questions

We have asked to rate 15 statements for analyzing a person's personality. These questions are based on 5 personality traits. Every trait has 3 questions each. The test consists of fifteen items that must rate on how true they are about you on a five-point scale where 1=Disagree, 3=Neutral, and 5=Agree. It takes most people 3-5 minutes to complete.

Extroversion	Neuroticism
● I don't talk a lot. ● I keep in the background. ● I don't mind being the center of attention.	● I get stressed out easily. ● I worry about things. ● I am easily disturbed.
Agreeable	**Conscientious**
● I feel little concern for others. ● I have a soft heart. ● I feel others' emotions.	● I leave my belongings around. ● I pay attention to details. ● I often forget to put things back in their proper place.
Personality	
● I have a rich vocabulary. ● I have a vivid imagination.	

• I have excellent ideas.	

3 Research Methods

Automatic Personality Classification system which uses Data Mining techniques to classify the personality of the applicants. The system uses algorithms like K-means, K-Nearest neighbor, Support vector machine and Big Five Model along with advanced data mining to mine user characteristics data and learn from the patterns. This learning can now be used to classify/predict user personality based on past classifications. The system analyses vast user characteristics and behaviors and based on the patterns observed, it stores its own user characteristics patterns in the model. The system now predicts new user personality based on personality data stored by classification of previous user data. This system is useful for predicting personality of applicants applying for various roles in an organization. We gave 15 statements to give rating on (1-5). After that all the ratings on every question are exported to csv file. We use K-means, K-Nearest neighbor, Support vector machine on these data.

4 Data

4.1 Data Collections

We collected our data from our friends. We provided them with our research statements and asked them how much they are agreeing with this statement. These statements will be provided by the google forum. After that, all the responses were exported to csv file. There are no null values in our data and all the numeric numbers are between 1 to 5.

	EXT1	EXT2	EXT3	EST1	EST2	EST3	AGR1	AGR2	AGR3	CSN1	CSN2	CSN3	OPN1	OPN2	OPN3
0	1	5	5	5	5	5	4	5	5	5	3	4	1	5	1
1	3	3	5	5	5	5	1	5	4	1	5	3	3	5	5
2	3	5	1	5	5	3	3	5	3	2	3	3	3	5	4
3	5	5	5	5	5	5	1	5	5	1	5	1	1	5	1
4	3	4	4	2	4	2	4	4	4	2	5	2	4	4	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
91	4	3	4	4	4	4	3	3	3	2	2	3	3	5	4
92	5	5	2	4	5	5	4	4	2	3	5	5	2	4	3
93	2	2	4	2	2	2	2	1	4	4	4	4	5	5	5
94	2	2	4	2	2	4	1	5	5	2	5	2	2	5	4
95	3	3	2	4	5	4	1	5	2	1	5	1	5	5	3

4.2 Data Analysis

We have a dataset where we ask to give ratting on our provided 15 statements. To classify a person's personality, we use different types of algorithms. K-means clustering is a very famous and powerful unsupervised machine learning algorithm. It is used to solve many complex unsupervised machine learning problems. We used this clustering method to identify the five types of personality. After finding all the personalities I use different types of algorithms. The system uses algorithms like K-means, K-Nearest neighbor, Support vector machine and Big Five Model along with advanced data mining to mine user characteristics data and learn from the patterns.

K-Means clustering. K-means clustering is a very famous and powerful unsupervised machine learning algorithm. It is used to solve many complex unsupervised machine learning problems. We used this clustering method to identify the five types of personality. After applying this clustering algorithm, we can identify the five types of personality persons.

K-Nearest neighbor. K-nearest neighbors (KNN) is a type of supervised learning algorithm. KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which is close to the test data. The KNN algorithm calculates the probability of the test data belonging to the classes of 'K' training data and class holds the highest probability will be selected. In our dataset there are 15 statements. Every participants give 15 ratings on etch statement.

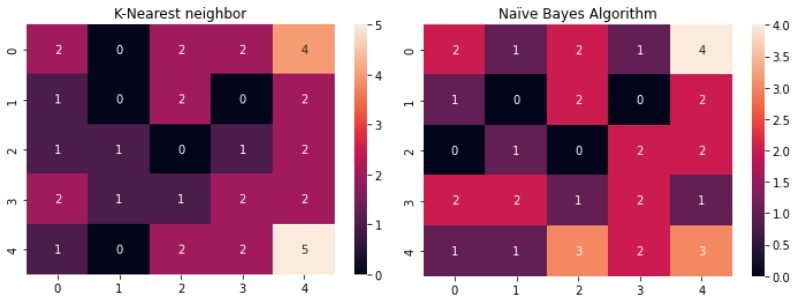
Naïve Bayes Algorithm. Naïve Bayes Algorithm, which is a type of an inductive learning algorithm, is considered to be one of the most efficient and effective algorithms that is widely used in data mining. The performance of Naïve Bayes Algorithm in classifying data is quite accurate because the conditional independence assumption on which the entire algorithm is set up is rarely true for the real world applications. The application of Bayes theorem forms the basis of Naïve Bayes Algorithm. A variation of Naïve Bayes Algorithm is Multinomial Naïve Bayes Algorithm which is also designed for classification purposes. The Multinomial Naïve Bayes Algorithm uses multinomial distribution in which it considers the either the number of times a particular word occurred or the weight of that particular word as a classification feature.

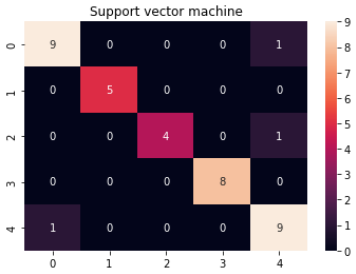
Support vector machine. Support Vector Machine which is a type of supervised learning algorithm analyses the data and recognizes pattern for classification purposes. A set of training data is taken and then it is marked as a part of category to predict whether the test document is a member of an existing class or not. The Support Vector Machine models represent the data set in the form of a point in space which is divided by either a line or hyperplane. The main idea behind the support vector machine algorithm is that if a classifier performs well at the most challenging comparisons, then it will definitely perform even better at the most easy comparisons.

5 Experimental Results

As we want to predict the personality by user statements ratings. After applying the clustering algorithms, we find the personality cluster data for every rating. After applying K-Means clustering we find this result. After applying the k-nearest neighbors' algorithm we find an accuracy of 81.57%. On the other hand, the Naïve Bayes Algorithm accuracy is 84.21%. Support vector machine we find the accuracy 92.10%.

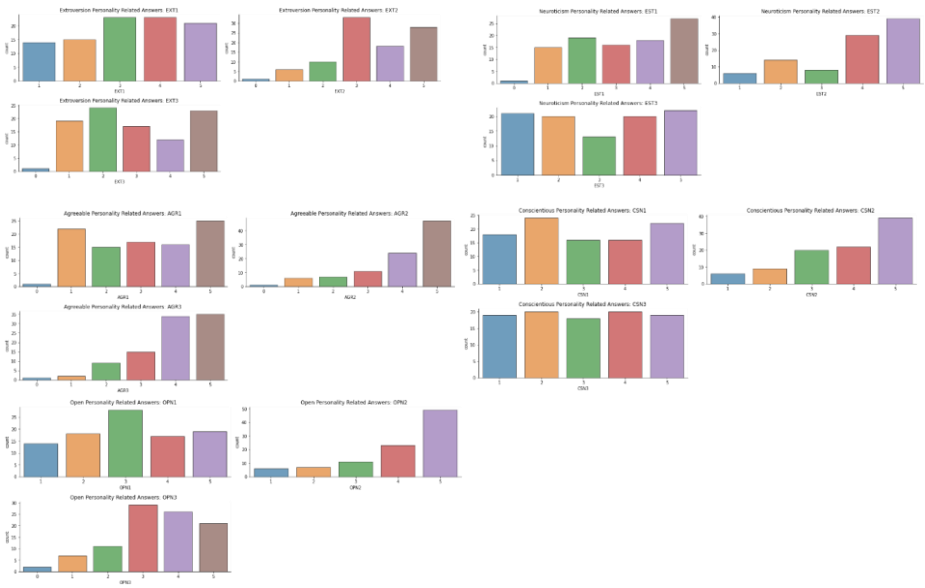
Confusion matrix.





6 Discussion

We collected data from our friends on google forums. Our data is about the rating (1-5) on 15 statements, So, it is already a cleaned data set. Here is the bar chart for five types of personality.



In these bar chart we can clearly see the rating of etch personality persons. Extroversion personality persons are not kept themself in the background and they don't mind being center of the attention. Narcotism personality person they get stressed out easily, worry about things, they are easily disturbed. If we see the bar chart, we can see that in our survey most of the people are not narcotism personalities. Now, if we look at the agreeable personality questions 5 ratings are higher.

In that case, most of the people in the survey are agreeable personalities. The conscientious Personality related question's results are a bit confusing. We cannot decide

from this result. In the open personality questions bar chat we can see that there are mutual of this personality.

In this section, we report and discuss how the classification algorithms performed for predicting the personality of the user. Support vector machine has the best accuracy in the three methods tested with an average accuracy of 70.89%. Support Vector Machine method performance was a little better than Naïve Bayes and K-Nearest neighbor due to the difficulties of separating a class of the dataset was quite accurate.

7 Conclusion

Personality analysis and prediction have increased very much in recent times. Extracting the personality of the user using the current system is very much helpful in various fields like the recruitment process, medical counseling, etc. Personality detection from a survey means extracting the behavior characteristics of the users taking the survey. We are providing a state-of-art review of an emerging field i.e., personality detection from the survey. Apart from the work done towards this system, future work mainly comprises the following objectives: For future work, we want to include more personality traits so that we can provide a more detailed personality to the user as well as predict personality using textual data and sentiment analysis. There can be a module where the user will be provided with career guidance and counseling sessions that matches his personality.

8 References

1. Asendorpf, J. B. (2001). Editorial: The Puzzle of Personality Types. European Journal of Personality, 16, S1-S5
2. Boehm, B., Asendorpf, J. B., & Avia, M. D. (2002). Replicable types and subtypes of personality: Spanish NEO-PI samples. European Journal of Personality, 16, 25–41
3. J. Golbeck, C. Robles, K.Turner(2011).Predicting personality with social media, In CHI'11 Extended Abstracts on Human.
4. Fazel Keshtkar, Candice Burkett, Haiying Li and Arthur C Graesser(2014).Using Data Mining Techniques to Detect the Personality of Players in an Educational Game, Springer International Publishing