

Personality classification using Data mining approach

Rajalaxmi Hegde¹, Sandeep Kumar Hegde², Sanjana³, Sapna Kotian⁴, Shreya C Shetty⁵

^{1,2,3,4} Department of Computer Science and Engineering, NMAMIT, India

Abstract—Personality classification refers to the psychological classification of different types of individual. This project deals with the areas where it determines the characteristics of a person. It can be helpful to classify person using Personality classification using data mining approach. In this paper, we aim to automate the personality prediction of the users by taking a personality test. The system uses classification algorithm i.e. N-closest neighbourhood algorithm (NCN). The analysis is done using vast set of data in data set and is been compared with the user input. This paper mainly focuses on classification algorithm.

Keywords—Personality classification, Data mining, NCN classifier.

I. INTRODUCTION

Personality is defined as the aspect with the set of perception, feeling and behavioural patterns that develop from botanical and external factors. Generally, there is no proper approval for definition of personality, mainly they focus on provocation and conceptual interactions. Even personality can be defined as traits that predict a person's behaviours. Personality identification was the old approach to identify the user's personality but now with the help of data mining techniques accuracy of this prediction has improved a way lot than old techniques.[1]

Data mining is the technique of finding pattern in huge data sets involving methods at the interaction of statistics, database systems and machine learning. Its overall goal is to produce information from datasets and transfer information.

The automated personality consists of comparing user's personality against standard personality tests taken. Mainly personality prediction depends on person's nature. Several tests will be taken by asking set of questions and depending on the answers chosen by the user, personality will be predicted. Classification algorithm used is N-closest neighbourhood. It is very important to process large volume of data and this can be done by Classification algorithm.[2]

The major goal of this paper is to give the outline for the growth of personality prediction depending upon the respective questions been answered. The outline of this paper is to predict personality of respective user and suitable career options.

II. LITERATURE SURVEY

This paper proposes how different multi-label classifiers like Binary Relevance, Classifier Chains and Random k-Label sets are been compared by the accuracy obtained from these models. How individual can be assigned more than one class label [3].

The paper aims at how through Social Media data can be accessed and the personality of an individual can be identified. It also states that the datasets which is been extracted from social media like facebook etc is relevant and efficiency of the algorithm can be improved [8].

Here in this paper they have specified how latest model Big Five Model Personality is used for Personality detection. They have attempted to build a system that can predict a person's personality based on Facebook user information [5].

This paper focuses on the accuracy of different learning algorithms like Naïve Bayes, Support Vector Machine and Decision tree. The accuracy is compared and the most efficient algorithm is been analyzed among the existing algorithms [1].

This paper provides an insight on development of different personality traits prediction through text on different social media platforms and how improvements can be made in existing techniques which can be applied in the future. And also, different kinds of existing datasets like twitter datasets, YouTube datasets, Facebook datasets etc [4].

Here in this reference paper they have majorly focused on the accuracy of different classification algorithms that are been used widely and this they estimated with the use of WEKA tool. The three algorithms are Decision tree, K-nearest neighbor and Naïve Bayes [2].

This paper deals how with the help of vast data mining algorithms we can predict a suitable career or profession for the students. However, this approach is beneficial for the academic progress of students [6].

This paper deals with the framework that's been designed by the authors which is effective at different domain. Datasets here are related to the comments, likes in Facebook account of user to predict the personality [7].

No.	Author	Data used for Processing	Tool Technique algorithm used	Merits	Demerits
1	Anisha yata, Prasanna Kante, T Sravani, B Malathi.	Textual data given by the user.	Binary Relevance, Classifier Chains and Random k-Label	Efficiency of base classifier KNN when used with multi label classifier gives the optimal result.[3]	None.
2	Anneke D.S. Rahmanto, Derwin suhartono, Williem and Veronica Ong.	Different Datasets available through social media like twitter, Facebook, YouTube etc.	Existing machine learning algorithms.	Personality prediction is done using text analysis on Social media.[4]	Large data sets which can be expensive.
3	Hendro, Rini Wongso, Derwin Suhartono, Yen Lina Prasetio, Tommy Tandra.	Statuses of the face book users and few manually collected users.	Comparing accuracy of Support Vector Machine (SVM), Naïve Bayes, Gradient Boosting Logistic Regression and Linear Discriminant Analysis (LDA).	Efficiency of classification algorithms are implemented using Social media related contents as their training datasets.	Restriction usage of large amount of datasets.
4	Sayali D. Jadhav, H. P. Channe	Weather and other factors to check the accuracy.	Here they find the accuracy of Naïve Bayes, KNN and Decision tree by Weka tool.	Estimating on the basis of accuracy the best classification algorithm to be Decision tree.[2]	However, the efficiency of algorithm is said to be depending on the applications and requirements.
5.	Avnish Kumar, Akshat Gawankar, Kunal Borge & Mr Nilesh M Patil.	Sample data related to student career options.	Data mining applications.	It helps to predict a suitable career options for students with effective and faster result prediction [6]	Doesn't give the specification of particular algorithms that can be used here.
6.	Manasi Ombhase, Prajakta Gogate, Tejas Patil, Karan Prof. Gayatri Hegde.	Sample data saved which is related to personality traits.	Naïve Bayes and Support vector Machine comparison is been done.	Personality prediction is done through text using advance machine learning algorithms.[1]	Efficiency between the advance algorithms not been specified.
7.	Candice Burkett, Haiying Li Arthur C. Graesser and Fazel Keshtkar.	One part of dataset deals with annotated containing personality excerpts based on Leary's Rose frameworks	Two methods are followed one deals with the two human judges experiment manually annotated the result and other method is approached by the usage of data mining approaches like Naïve Bayes, SVM, J48.	N gram is the best specific detection technique that's been estimated accordingly.[7]	No much description the algorithms for sentiment analysis.
8.	Janhavi Pednekar, Shraddha Dubey.	Data sets related to the social media contents.	Techniques used for sentiment analysis are Rule Based Classifier (RBC), General Inquirer based classifier (GIBC), Induction Rule Based Classifier (IRBC) and Statistics Based Classifier (SBC),	Taking the Likes, comments from the Facebook as datasets which can be used to predict the nature of user.[8]	No much description about other classifiers are been mentioned.

III. PROPOSED SYSTEM

The system design here includes the Information Retrieval, Data Set, Classification Algorithms, Personality Trait Repository and Personality Trait Report. Initially the user has to response to the given set of questions and those responses are been collected, Later, these responses are compared with the already existing training data sets. However, this comparison are performed using Classification Algorithm i.e. N-closest neighbourhood.

Personality trait repository basically stores all the personalities like Extroversion, Introversion, Sensitive etc. So once the processing and is completed the result i.e. the predicted personality of particular user is displayed on the screen. (fig 1.1)

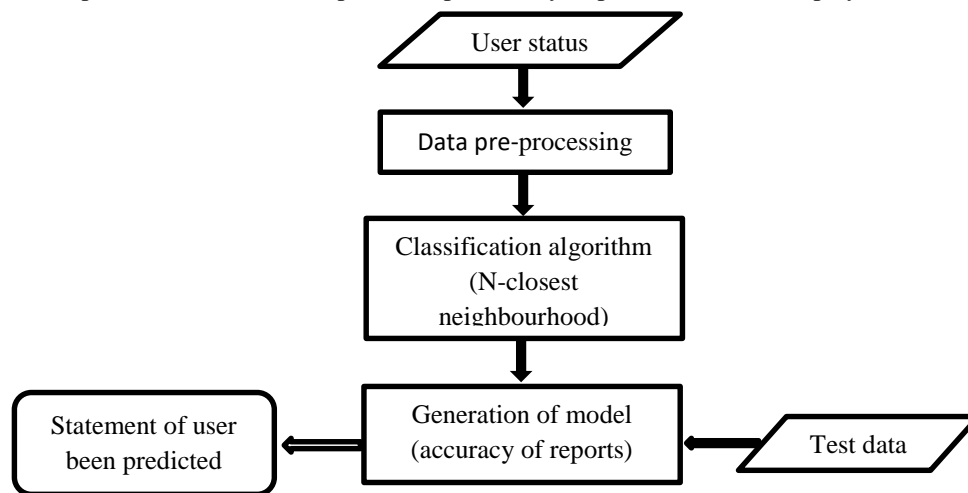


Figure 1.1 Flowchart

IV. IMPLEMENTATION

A. Datasets

Data set mainly consists of single statistical data matrix, in which every column represents a specific variable and the row represents the possible combinations of answers for the questions. In this project, the dataset consists of values or responses answered by the user for the given set of questions, user personality, best career option. The responses are compared with the already existing training set. User personality such as Extraversion(E) or Introversion(I), Sensing(S) or Intuitive(I), Feeling(F) or Thinking(T), Judging(J) or Perceiving(P) are taken. With the help of these personalities the best career for each personality can be predicted, i.e. if the person has a combination of ESTJ can be a chef.

B. Data pre-processing

All the information in English go through pre-processing level before getting processed. Pre-processing is used to remove all the lower case, symbols, names, spaces etc. for example any word goes through pre-processing stage and after this word will be processed and converted into English.

C. N-closest neighbourhood Classification

This is the simplest algorithm among other algorithms in machine learning which is easy to understand and implement. Main principle used is the pattern with similar features which always lie in close sector [9]. k-nearest is example of learning algorithm. Classifiers based on this example are called lazy learners which stores all the training sets and classifiers are aren't built until new or unlabelled sets need to be classified [10]. Lazy-learning algorithm requires less computation time during the training process but more time during the classification process compared to fast-learning algorithm (i.e. decision tree, neural network and bayes network [11][12].

By analysing and understanding the above algorithm we have estimated the new algorithm that is N-closest neighbourhood algorithm (NCN). Flow of the diagram is given in fig 1.2.

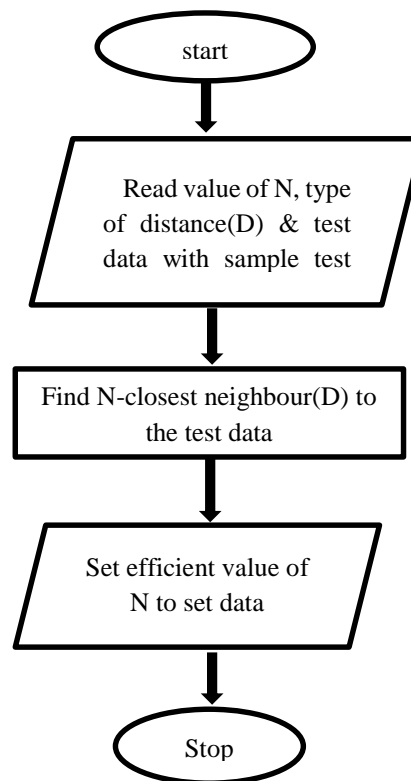


Figure 1.2 Flowchart of NCN classifier

N-closest neighbour majorly depends upon the selection of value for N. Basically in this we have to compare the sample test (i.e. X) with the already existing training sets. So, we use X sample test to be classified, here N closest neighbours are estimated and then X is allocated to a particular label of class depending upon the majority neighbours are attached to specific label. Selection of N value is very important, so it is essential to make sure that N value is as small as possible so the accuracy can be assured, if the N value is large then there is possibility that N closest neighbours can misclassify the X sample test with some other class labels. The distance has to be estimated from the sample test to the closest neighbouring points by Euclidean distance formula. [13]

The steps of NCN classifier are as follows:

1. Estimate N.
2. Evaluate distance between sample input and training samples.
3. Categorise the distances.
4. Clasp highest N-closest neighbour.
5. Select the efficient and the nearest value.
6. Identify class label for input sample with more neighbour.

Euclidean distance is calculated as:

If there are points (e1, f1) and (e2, f2) in 2- dimensional space, then Euclidean distance between them is

$$x = \sqrt{(e2 - e1)^2 + (f2 - f1)^2}$$

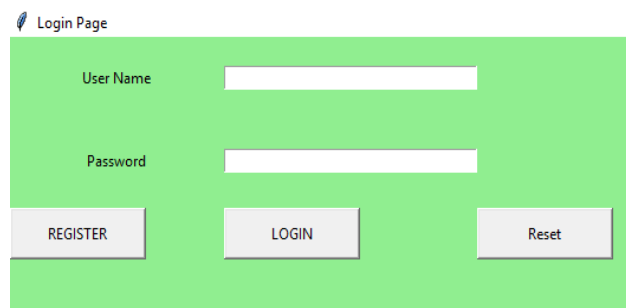
It is used to find the distance between the sample test and closest distance.

D. Generation of model

The responses are reserved according to the personality surveys that is been conducted and the algorithm is used to group the reaction into collection of non-covered personality types that occur across all survey datasets with disproportionate frequency and it will track most average personality type. The result will be accurate about the user's personality depending on the questions been answered in the survey. Main output will be the personality of a user and suitable career option.

V. RESULT

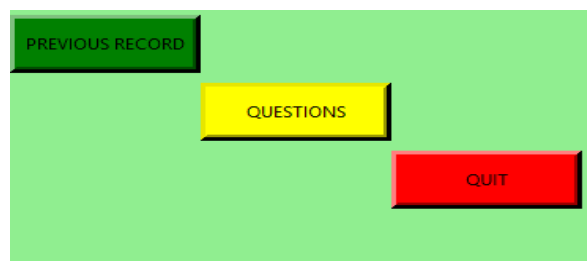
This paper gives an insight on existing attempts of the task of personality prediction and suitable career option from questions been answered in a survey. Firstly, the user has to register by filling the respective information and then the user can login whenever he wants to (fig 1.3).



The image shows a login page with a light green background. At the top left, there is a small icon of a feather and the text 'Login Page'. Below this, there are two input fields: 'User Name' and 'Password'. At the bottom, there are three buttons: 'REGISTER', 'LOGIN', and 'Reset'.

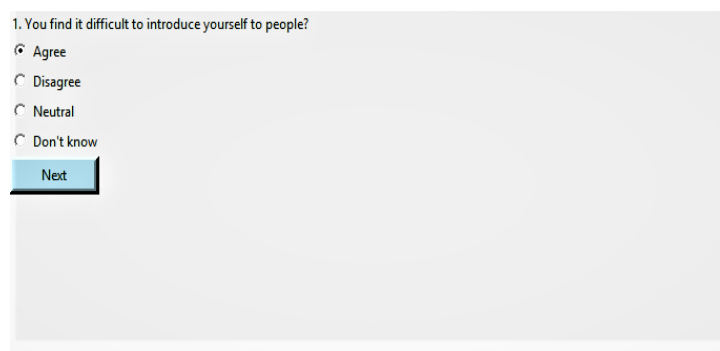
Figure 1.3 Login page

When the user logs in, he can take the personality test by answering the set of questions and he can even view the previous records whenever he wants to login. (fig 1.4 and fig 1.5)



The image shows a main menu with a light green background. There are three buttons: 'PREVIOUS RECORD' (green), 'QUESTIONS' (yellow), and 'QUIT' (red).

Figure 1.4



The image shows a survey page with a light gray background. The question is '1. You find it difficult to introduce yourself to people?'. There are four radio button options: 'Agree', 'Disagree', 'Neutral', and 'Don't know'. Below the options is a 'Next' button.

Figure 1.5 Survey page

After the personality test is taken, the user's personality and best career option is been predicted. (fig 1.6)



The image shows an output page with a light green background. The text reads: 'YOUR BEHAVIOUR BELONGS introversion sensing thinking !!' and '-Suggested profession is reporter or engineer'.

Figure 1.6 Output page

VI. CONCLUSION

Further improvements for personality prediction can be made by applying new languages and more worthy algorithm or processing methods to attain higher perfection(accuracy). The results can show that machine learning can improve the accuracy even if the accuracy level is quite low for few traits. It is due to small number of dataset used. However, it shows that traditional machine learning and deep learning can perform the results of previous studies using the same dataset.

Further scope, we want to include proposed method in current research to sentiment analysis, opinion mining, as well as detection of emotions in other domains. Also, we want to extend the method in this work to apply in Big-Five personality detection.

VII. REFERENCES

- [1] Manasi Ombhase, Student, PCE, Prajakta Gogate, Student, PCE, Tejas Patil, Student, PCE, Karan Nair, Student, PCE and Prof. Gayatri Hegde, Faculty, PCE, "Automated Personality Classification Using Data Mining Techniques"
- [2] Sayali D. Jadhav¹, H. P. Channe² "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques", Department of Computer Engineering, Pune Institute of Computer Technology, Savitribai Phule Pune University, Pune, India
- [3] Anisha Yata¹, Prasanna Kante², T Sravani³, B Malathi⁴, "Personality Recognition using Multi-Label Classification" 2018.
- [4] Veronica Ong, Anneke D. S. Rahmanto, Williem and Derwin Suhartono, "Exploring Personality Prediction from Text on Social Media": A Literature Review 2017.
- [5] Tommy Tandra, Hendro, Derwin Suhartono*, Rini Wongso, and Yen Lina Prasetio "Personality Prediction System from Facebook Users" Computer Science Department, School of Computer Science, Bina Nusantara University, Jl. K. H. Syahdan No. 9 Kemanggis, Jakarta 11480, Indonesia
- [6] Avnish Kumar¹, Akshat Gawankar², Kunal Borge³ & Mr Nilesh M Patil⁴ .1 2 3 B.E IT Student, "Student Profile & Personality Prediction using Data Mining Algorithms" Information Technology, Rajiv Gandhi Institute of Technology, Mumbai, India 4 Assistant Professor, Information Technology, Rajiv Gandhi Institute of Technology, Maharashtra, India
- [7] Fazel Keshtkar, Candice Burkett, Haiying Li and Arthur C. Graesser, "Using Data Mining Techniques to Detect the Personality of Players in an Educational Game".
- [8] Janhavi Pednekar¹, Shraddha Dubey² 1,2 Symbiosis, "Identifying Personality Trait using Social Media": A Data Mining Approach Institute of Computer Studies and Research, Symbiosis International University, {janhavi. pednekar, shraddha.dubey}@sicsr.ac.in
- [9] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification", IEEE Transactions on Information Theory, vol. 13, No. 1, pp. 21-27, 1967.
- [10] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Elsevier, 2011.
- [11] K. P. Soman, "Insight into Data Mining Theory and Practice", New Delhi: PHI, 2006.
- [12] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica, vol. 31, pp. 249-268, 2007.
- [13] Bhavesh Patankar and Dr. Vijay Chavda, "A Comparative Study of Decision Tree, Naive Bayesian and k-nn Classifiers in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 12, December 2014.