# Personality Traits Identification using Rough sets based Machine Learning

Umang Gupta
Department of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi - 110016
Email: umang.18sep@gmail.com

Niladri Chatterjee
Department of Mathematics
Indian Institute of Technology Delhi
New Delhi - 110016
Email: niladri@maths.iitd.ac.in

***Abstract*—Prediction of human behavior from his/her traits has long been sought by cognitive scientists. Human traits are often embedded in one's writings. Although some work has been done on identification of traits from essays, very little work can be found on extracting personality traits from written texts. Psychological studies suggest that extraction and prediction of rules from a data has been long pursued, and several methods have been proposed. In the present work we used *Rough sets* to extract the rules for prediction of personality traits. Rough Set is a comparatively recent method that has been effective in various fields such as medical, geological and other fields where intelligent decision making is required. Our experiments with rough sets in predicting personality traits produced encouraging results.**

***Keywords-component; personality recongnition, big-five model of personality, rough sets, machine learning***

## I. INTRODUCTION

Prediction of human behavior is a classical problem of psychology. Psychologists in general believe that one's personality affects various aspects of his/her behavior, such as, job performance, effectiveness, dominance [1]. Recent studies are focused on automatic identification of traits (such as emotion, sentiment, dominance, deception) from the conversation and texts. Very little work, however, has been done on extracting personality traits from written texts. In this respect, our focus is on identification of Big five traits for a person from his/her written text. Big Five model of personality is a standard model for personality traits [2]. According to Big Five model, personality is assessed along five dimensions –

a. *Extroversion* **(**sociable, assertive and playful**)**
b. *Neuroticism* **(**insecure and anxious**)**
c. *Agreeableness* **(**friendly, trustable and co-operative**)**
d. *Conscientiousness* **(**organized**,** sincere**)**
e. *Openness to experience or openness* **(**intellect, good learner**)**

Psychological experiments have successfully demonstrated the importance of Big Five traits in identification of human-behavior related traits, such as, deception, job performance, and other aspects. For example [3]:

- Extroverts are better at deception,
- People with openness to experience tend to be more successful at work places,
- An agreeable person is good at deception but he will seldom try to lie.

Psychologists have derived certain thumb rules to identify the traits looking at the *utterance*s of the person.[1] These rules can be exploited in an automatic personality identification system. In present work, we trained the system to identify these personality traits from the text datasets, which were tagged for the personality score.

Automatic prediction of personality has variety of applications, such as, identification of right candidate for the job [4], analyzing deception [3], and identifying right matches from the chat text on dating and other social websites [5]. It will be helpful in making user-friendly systems that will identify the person's behavioral traits automatically. The present work focuses on identification of Big Five Personality traits from written set of texts tagged for the same purpose. Information acquired from the texts is purely linguistic, and no reference to the person is made.

We used rough set theory [6], [7] to make predictions and classifications. Rough sets are well-known tools for knowledge discovery from data. Rough sets are being used for various processes, namely – attribute selection, data reduction, rule generation and pattern extraction [8], [9]. Rough set based techniques have emerged as an efficient machine learning technique for extracting knowledge from incomplete datasets. Rough set theory assumes that given the datasets, (i.e. features/ condition attributes), examples can be classified into the decision classes; and the accuracy can be evaluated. A unique feature of Rough sets is that here the information is neither exaggerated/ or approximated, nor is it deprecated; rather it is used as it is to evaluate the approximations. We have used Rough Set based software [10] to extract rules from the feature set to train the system based on observed data, and test the overall scheme of classification on test data.

The paper is organized as follows. In Section II we describe rough sets in detail. Section III discusses data and pre-processing, Section IV has results. Section V concludes the paper with possible future directions.

[1] Utterances stand for both spoken as well as written words. In the present work we consider utterances for the writings only.

978-0-7695-5066-4/13 $26.00 © 2013 IEEE
DOI 10.1109/ISCBI.2013.44

## II. ROUGH SETS

### A. Basic concepts

Rough sets have emerged as a mathematical tool for modeling imperfect knowledge. Rough sets are good for evaluating significance of the data from information system. We define Information system $S$ as:

$$S = (U, C, D, f)$$

Where $U$ is the set of objects, $C$ is set of condition attributes, $D$ is set of decision attributes and $f$ is a relation such that

$$f: U \rightarrow Va$$

such that for any a ϵ A, $V_a$ is set of its values. For each attribute a ϵ A and x ϵ U.

$$f(x, a) = Va$$

It is basically a data table consisting of decision and conditions. We consider all the examples/objects to be described by some attributes/features. The main concept behind rough set is indiscernibility relation.

### B. Indiscernibility relation

With respect to f and A, the universe U is divided into partition such that – all partitions are disjoint and Two elements belong to same partition if

$$f(x, a) = f(y, a) \forall\ a \epsilon A$$

*x*, *y* are said to be indiscernible and the set of all indiscernible elements is denoted as IND(A).

These partitions represent the smallest granule of knowledge. It is the maximum information that can be discerned about the system.

### C. Approximations

Given a subset P, such that

$$B \subseteq A,$$

We can define IND (B) and all the elements belonging to a particular partition are called B-indiscernible. Also given a subset Q⊆U, we define lower and upper approximations as:

$$lower(A, Y) = \{X: X \epsilon\ U/IND(A), X \subseteq Y\}$$
$$upper(A, Y) = \{X: X \epsilon\ U/IND(A), X \cup Y \neq \Phi\}$$

*Lower approximation* is interpreted as the objects that surely belong to a particular partition. This means they can be classified with 100% confidence; whereas *upper approximation* is interpreted as the set of all the objects that can be classified to a particular class with some confidence - but not necessarily with 100% confidence.

Similarly *positive region* is defined as the union of all the lower approximations of P given X, such that

$$X \subseteq IND(Q).$$

To say, all the elements belonging to positive region can be classified with 100% confidence.

$$POS(P, Q) = \bigcup_{X \epsilon Q} lower(P, X)$$

*Accuracy of approximation* is defined as the ratio of cardinality of lower and upper approximations. Difference between upper and lower approximations is called *boundary of approximation*. It is precisely the objects about which decision cannot be made with full confidence.

### D. Dependency of attributes

We say that an attribute D depends on C if all values of D can be uniquely determined by C. Hence D depends totally on C. However, it is possible that C may not completely describe D but partially determine D; hence we define partial dependency of attributes. A coefficient of dependency, k is used to describe the partial dependency (0 < k <1)

$$k = \gamma(C, D) = |POS(C, D)|/|U|$$

### E. Reduction of attributes

Attribute selection is a key problem. Rough set provide a good method to eliminate the superfluous data. We say an attribute can be dispensed from the set of attributes as irrelevant if

$$IND(B) = IND(B - \{a\})$$

Then attribute 'a' is dispensable. So, one can easily remove the superfluous data. The attributes left are called *reduct of A* and denoted as red (A). Intersection of all the reducts is called the *core of A* and denoted as core (A). When we have condition and decision attribute this condition reduces to preservation of positive region.

$$POS(C, D) = POS(C - \{a\}, D))$$

Also significance of each attribute can be defined as

$$\sigma(C, D, B) = [\gamma(C, D) - \gamma(C - B, D)]/\gamma(C, D)$$

Where is $\sigma(C, D, B)$ the significance of attribute B with respect to condition attributes C and decision attribute D.

### F. Other concepts

*Support, Coverage and Certainty*: Let the rule be

$$C \rightarrow D: \{C_1, C_2 \ldots C_n\} \rightarrow \{D_1, D_{2 \ldots} D_n\}$$

Support of decision rule is defined as

$$supp(C, D) = |C(x) \cap D(x)|/|U|$$

This implies the strength of decision.

On the same line certainty factor is defined as the probability that the rule predicts the right decision.

$$cer(C, D) = |C(x) \cap D(x)|/|C(x)|$$

And coverage of a rule is defined as probability, given the decision from D, what is the importance of condition.

$$cov(C, D) = |C(x) \cap D(x)|/|D(x)|$$

Certainty factor can be used to evaluate the accuracy of decision and coverage factor gives the reason for the decision. Hence rough sets give a reason for the decisions.

Software based implementation of rough set is described in [10]. We used ROSE2 software for our work. Some other implementations are described in [8], [9] and [11].

## III. DATASET AND PRE-PROCESSING

We used text files as datasets. This was the same data as in [1]. Data was provided by Prof. James Pennebaker, Department of Psychology, University of Texas. These files were tagged for the Big Five Personality Traits. These files were created by asking students to write whatever comes to their mind for 20 minutes. Their personality was assessed by a personality questionnaire. Hence we had text files along with tags for the Big Five traits. This data set had to be processed to extract the relevant features and then the features have to written in a particular format to work with ROSE2 Software.

Sample clippings of texts:

*"Matter of fact I can't even have a real stream of consciousness that I would usually have because I am doing this. I am too busy thinking about how this thing works and analyzing within myself what type of thoughts should be going through my head. I don't know how to word what I mean really because I don't have time to think about it. I wonder if you are really going to read this. do you really find this exciting. I wonder if whoever is reading this will think I'm a boring person because I'm not talking about nothing. Or I wonder if they will wonder how I made it to UT because you would think I couldn't spell or be eloquent by looking at this. I don't know I Love the Lord. someone just walked in and broke my stream. They were talking to me. Dang I was flowing. O well I guess that's what this is for."*

The original texts are of about 800 words. The scores for the 5 traits as evaluated from the questionnaire are given as follows:

Openness - 22.00, Consciousness - 27.00, Extroversion - 27.00, Agreeableness - 24.00, Neuroticism - 4.00.

The scores were obtained by evaluating standard personality questionnaires. We have normalized these scores by global means to make them binary in the following way:

Features were normalized against the mean and the ones more than the average were considered to be 1 else 0. This conversion helps us in Rough set based calculations in a straightforward way.

## IV. EXPERIMENTS AND RESULTS

### A. Extracting the features:

We used linguistic features as condition attribute as given in [1]. Features were extracted using two sources:

- LIWC (Linguistic Inquiry and Word Count) dictionary [13]. It contains more than 5000 words tagged under different sections as linguistic words (e.g. pronoun, adjective, noun etc.), psychological process (e.g. emotions), relativity (e.g. time, past, future), and personal process (e.g. home, family, job). LIWC dictionary has words tagged under different categories like anger words, positive emotions, negative emotions, anxiety words etc.
- NLTK [12], NLP toolbox of python programing language. It tagged the words by POS (parts of speech) tagging for extracting grammatical features (e.g. noun, pronoun).

The extracted features have been classified into 89 categories, such as *anger* words, *we* words, *numerals*. Features like *anger* words, *we* words, emotion words were extracted using LIWC dictionary [13] and grammatical features (like articles, pronouns, numerals) were extracted using NLTK [12] by POS (parts of speech) tagging. The frequency of words under particular category has been considered as the value of the feature in a particular text.

### B. Validation and training:

We run LEM [13] algorithm on this feature set using ROSE2. ROSE2 software has two validation methods - basic minimal covering (LEM2) [8] and satisfactory description. LEM2 algorithm gives the minimal set of rules.

In our experiments we used both the methods and evaluated the result. Table 1 provides the results of using all the features together.

TABLE I. BASIC MINIMAL COVERING (LEM2 ALGORITHM), 10FOLD VALIDATION

| Trait | Accuracy |
|---|---|
| Openness | 83.07±2.92 |
| Extrovert | 63.07±6.2 |
| Consciousness | 77.73±5.1 |
| Neuroticism | 54.53±3.84 |
| Agreeableness | 84.67±3.06 |

The values in Table are written as *Mean* ± *Standard Deviation* in order to give an understanding of the central tendency and dispersion of the obtained results.

TABLE II. COMPARISON WITH RESULTS IN [1]

| Traits | Accuracies as reported in [1] | Calculated accuracies (all features) |
|---|---|---|
| Extraversion | 54.93 | 63.0 ± 6.2 |
| Conscientiousness | 55.29 | 77.73 ± 5.1 |
| Openness | 62.11 | 83.07 ± 2.92 |
| Agreeableness | 55.78 | 84.67 ± 3.06 |
| Neuroticism | 57.35 | 54.53 ± 3.84 |

Table II provides a comparison of our obtained results of accuracy with the values obtained by François Mairesse et al, [1]. It is evident from Table II that the accuracy for *openness*, *consciousness* and *agreeableness* are far better than those in [1] which used a Support Vector Machine based implementation which gave the best accuracy. However, such improvements are not noticeable for *neuroticism* and *extroversion*. The reason for not getting

good results for extroversion and neuroticism can be explained in the following way:

The number of rules inducted for both of them is large (>80), which implies that these two traits lack definite patterns. As a consequence, more rules are generated as the Rough set based scheme takes care of different cases existing in our training database. This in turn renders the classification poor. This problem might have been induced due to binary discretization of features as was explained in Section III. Our scheme of discretization is rather simplistic. In future we plan to work on multiclass discretization of the feature values.

Also to understand the quantitative variation in learning with data sets, learning was done by using different number of training examples. Table III shows the result. In these experiments we have split the entire dataset into training and test datasets in different proportions. As expected the accuracy of prediction increases with the increase in the size training data. However, here too one can notice that for neuroticism and extroversion improvement is marginal with the increase in the training data size. This also shows that modeling for these decisions is not an easy task.

This gives motivation to further explore the Rough set methods in this regards.

## V. CONCLUSION

This paper discusses and interesting application of Rough sets for the task of personality classification. Sufficient improvement in accuracies is found over what has been achieved using SVM based approach. The advantage of the proposed scheme comes two reasons as proposed in our scheme:

- Identification of the right sets of features. Our scheme of using both LIWC features and linguistic features enabled us to identify a comprehensive set of condition features which in turn helped in better trait classification of human beings.
- Use of Rough sets to distill rules from the raw data. However, in our opinion, trivial binarization of the features made the scheme less effective. And we feel that the results can be further improved with multi class discretization.

This provides a motivation to work further on rough sets based algorithm for personality (and other related) classifications. There are many complex variants of Rough theory (e.g. variable precision, fuzzy rough sets) that are expected to perform better than the basic Rough set model used by us. In future, we aim to use these schemes for personality classification.

## ACKNOWLEDGMENT

We would like to acknowledge Prof. James W. Pennebaker for providing the tagged datasets.

## REFERENCES

[1]. François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Int. Res.* 30, 1 (November 2007), 457-500.

[2]. An alternative "description of personality": The Big-Five factor structure. Goldberg, Lewis R. Journal of Personality and Social Psychology, Vol 59(6), Dec 1990, 1216-1229.

[3]. Heinrich, C. U. and Borkenau, P. (1998), Deception and Deception Detection: The Role of Cross-Modal Inconsistency. Journal of Personality, 66: 687–712.

[4]. Chen-Fu Chien, Li-Fei Chen, "Using Rough Set Theory to Recruit and Retain High-Potential Talents for Semiconductor Manufacturing", IEEE Transactions on Semiconductor Manufacturing, vol 20, no. 4 Nov. 2007.

[5]. M.Brent Donnellan, Rand D. Conger, Chalandra M. Bryant, The Big Five and enduring marriages, Journal of Research in Personality, Volume 38, Issue 5, October 2004, Pages 481-504

[6]. Pawlak, Zdzisław. "Rough sets." *International Journal of Parallel Programming*11.5 (1982): 341-356.

[7]. Z Pawlak, "Rough sets", lectures at Tarragona University seminar on formal languages and rough sets in 2003

[8]. Stefanowski, Jerzy. "On rough set based approaches to induction of decision rules." *Rough sets in knowledge discovery* 1.1 (1998): 500-529.

[9]. Aboul Ella Hassanien and Jafar M. H. Ali. 2004. Rough Set Approach for Generation of Classification Rules of Breast Cancer Data. *Informatica* 15, 1 (January 2004), 23-38.

[10]. B.Predki, R.Slowinski, J.Stefanowski, R.Susmaga, Sz.Wilk: "ROSE - Software Implementation of the Rough Set Theory", Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence, vol. 1424. Springer-Verlag, Berlin (1998).

[11]. Chmielewski, Michal R., et al. "The rule induction system LERS-a version for personal computers." *Foundations of Computing and Decision Sciences* 18.3-4 (1993): 181-212.

[12]. Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc. ISBN 0-596-51649-5.

[13]. "LIWC: Linguistic Inquiry and Word Count" Internet: http://www.liwc.net/ [Oct 18,2012]

TABLE III. VARIAITION IN ACCURACIES WITH THE DATA

| % of training data | 20% | 40% | 60% | 80% | 100% |
|---|---|---|---|---|---|
| Openness | 65.81 ± 8.68 | 66.98 ± 6.09 | 75.60 ± 5.01 | 82.00 ± 2.96 | 85.45 ± 4.07 |
| Consciousness | 66.67 ± 7.96 | 67.39 ± 5.16 | 75.85 ± 5.63 | 76.83 ± 3.61 | 78.50 ± 3.01 |
| Extroversion | 59.10 ± 8.88 | 59.92 ± 7.88 | 63.08 ± 4.06 | 62.67 ± 3.67 | 62.22 ± 6.76 |
| Agreeableness | 73.00 ± 10.26 | 81.46 ± -7.34 | 86.38 ± 5.70 | 86.83 ± 3.29 | 86.12 ± 3.15 |
| Neuroticism | 49.05 ± 11.29 | 52.93 ± 9.40 | 53.17 ± 8.11 | 53.33 ± 6.10 | 53.41 ± 4.11 |

Note: The accuracies may vary slightly on differnet runs, hence error boxes are quoted along with the actual accuracies.