Evaluating Effectiveness of Classification Algorithms on Personality Prediction Dataset

Suril Shah
Computer Engineering Dept.
Dwarkadas J. Sanghvi COE
Vile Parle, Mumbai, India

Sagar Vikmani
Computer Engineering Dept.
Dwarkadas J. Sanghvi COE
Vile Parle, Mumbai, India

Sahil Modak
Computer Engineering Dept.
Dwarkadas J. Sanghvi COE
Vile Parle, Mumbai, India

Prof. Kiran Bhowmick
Computer Engineering Dept.
Dwarkadas J. Sanghvi COE
Vile Parle, Mumbai, India

***Abstract*—Classification in machine learning, refers to the process of categorizing given input data pieces into certain given groups. There are different types of classification algorithms that are widely used based on bayes, trees, functions or rules. The competence of these algorithmic methods has been a major issue since a long time and has caught the interests of a large researching community. In this paper we study the effectiveness of Rule-Based classifiers. There are several algorithms for rule classifier including Ridor, DTNB, JRip, OneR, NNge, ZeroR and many more. This paper presents a comparative analysis of Decision Table and Conjunctive Rule, two different classification algorithm, to classify and predict personality based on the Big Five Model dataset and supports the same with implementation results on WEKA.**

***Keywords*— Data mining, Big Five Model, Personality prediction, Classification, Decision table, Conjunctive Rule, Comparative Analysis**

I. INTRODUCTION

The process of discovering meaningful, new and interesting correlation, trends and patterns by sifting through large amounts of data, by using pattern recognition technologies as well as statistical and mathematical technique is referred to as Data mining. Data mining comprises of more than gathering and handling data; it also encompasses data prediction and analysis. Mistakes often happen while people try to examine or establish relationships between various features, making it problematic to find solutions to given problems. Machine learning can prove to be beneficent while solving these problems, refining the effectiveness of systems and the machines' designs. This paper particularly is concerned with the classification problems [1].

In Machine learning, classification problem can be viewed as an algorithmic practice for distributing given data into one among the given categories. A Classifier is referred to as an algorithm that implements classification. The input data can be referred to as an instance and the categories as classes. The features of the instances can be labeled by a vector of characteristics. These characteristics could be ordinal, nominal, real or integer valued. Various data mining algorithms work in terms of categorical data only, requiring the real or integer valued data to be changed into groups. Classification, a supervised technique, learns to classify new instances centered on the knowledge gained from training instances' sets. Clustering, a similar unsupervised procedure, too groups the input on the basis of innate resemblance measures.

Classification and Clustering, illustrations of general pattern recognition problems, both assign certain output values to a given input values. Classification schemes in machine learning brought from observed data are first rated by their predictive precision. The transparency of a classifier is frequently significant as well in practice. Hence, rule-based classifiers are more popular, since rules can be rather easily interpreted by humans.

The main aim of the paper is to study the performance of two of the classification algorithms. The remaining paper is organized into 7 sections. Section II gives an overview of the rule based classification algorithms used in this paper. The next section presents an overview of the Big Five Model of personality prediction. Section IV summarizes the different performance evaluation measures for the classifiers. The section V and VI deal with the dataset overview and the empirical results. Conclusions are drawn from the experimental results in section VII followed by the references in the next section.

II. RULE BASED CLASSIFIER ALGORITHMS

A. Decision Table

Definition: First Decision table for data set S with n attributes B_1, B_2, B_3, B_4 ..., B_n is a table with schema R (B_1, B_2, B_3, B_4 ..., B_n, class, sup, conf). A row R_i = (b_{1i} , b_{2i} , b_{3i}, b_{4i} ..., b_{ni}, e_i, sup_i, $conf_i$) in table R characterizes a classification rule, where b_{ij} ($1 \leq j \leq n$) can be either from DOM(B_i) or a special value ANY, $e_i \in \{e_1, e_2, ..., e_m\}$, $minsup \leq sup_i \leq 1$, and $minconf \leq conf_i \leq 1$ and the thresholds -minsup and minconf -are predetermined. The rule is inferred as: if (B_1 = b_1) and (B_2 = b_2) and … and (B_n= b_n) then class = e_i having probability $conf_i$ and support sup_i, where $b_j \neq$ ANY, $1 \leq j \leq n$[6].

A decision table is composed of two components:
(i) A list of attributes which is known as a schema
(ii) A multi-set of labeled instances known as the body.

There is a corresponding value for each of the attributes in the schema as well as for the label. The set of instances having the same values for a given schema attributes are collectively known as a cell. The structure of the decision table is similar to a relational table, where every row holds the mean of all the records for each possible combination of the attributes. A hierarchy of tables is constructed, after loading the decision table into the memory, where every new table is one level higher in the hierarchy and has two attributes less than the previous table. Finally, the table at the top-most level contains a single row which represents all the data. In addition to column for each attribute, there is also a column for the record count, and a column representing a vector of probabilities.

For example, consider Table 2, a decision table constructed from the given data where minconf = 1.0 and minsup = 0.20

TABLE I. DATA GIVEN

age-group	car-type	risk
young	family	high
young	sport	high
middle	sport	high
old	family	low
middle	family	low

TABLE II. DECISION TABLE

age-group	car-type	risk	sup	conf
Young	ANY	High	0.40	1.00
ANY	Sport	High	0.40	1.00
old	ANY	Low	0.20	1.00
middle	family	low	0.20	1.00

The value of 'conf' represents the conditional probability of a tuple having the designated class label given the values of its attribute. The support of the rule characterized by the row is given by the 'sup' column.

The table is referred to as a decision table, since each row in the table denotes a rule which can be used to govern the class of a sample with given values of a certain attribute. In the above example, the following rules are entailed by the table:

- *age-group* = "Young" → risk = "high" (40%, 100%)
- *age-group*= "old" → risk = "low" (40%, 100%)
- *car-type*= "Sport" → risk = "high" (20%, 100%)
- (age-group ="middle"), (car-type = "family")
→ risk = "low" (20%, 100%)

Decision Table Classifier:

Decision Table classifier [6] algorithm is used to summarize the dataset by using a decision table containing the same number of attributes as that of the original dataset. A new data item is allocated a category by searching the line in the decision table that is equivalent to the values contained in the non-class of the data item. Wrapper method is used by the decision table classifier algorithm to find a considerable subset of attributes to be included in the table. By eliminating attributes that that have little or no contribution to a test model, the decision table classifier algorithm minimizes the possibility of over-fitting and constructs a much smaller and condensed decision table. Greedy approach is employed for searching the attribute space, either a top to bottom approach or bottom to top. A top-to-bottom approach adds attributes at each stage. This is also known as forward selection. A bottom-to-top approach is initiated with a full set of attributes and deletion of attributes takes place one at a time. Hence, this approach is also known as backward elimination.

There are two variants of decision table classifiers. The first decision table classifier is called Decision Table Majority (DTMaj). This classifier returns the major part of the training set if the cell of the decision table which matches the new instance is empty. The second classifier is known as Decision Table Local (DTLoc). If the matching cell is empty, this classifier searches for an entry in the decision table with fewer matching attributes. This DTLoc therefore returns an answer from the native region.

B. *Conjunctive Rule*

Conjunctive Rule[3] is a rule for decision-making where in, the expected buyer allots smallest values for a number of factors and rejects any result not meeting the lowest threshold value on all of the factors i.e. a better performance on one factor cannot remunerate for insufficiency on another. Conjunctive rule uses the relation of AND logical to link stimulus attributes.

The algorithm of ConjunctiveRule applies a single conjunctive rule learner which is able to predict for nominal as well as numeric class labels. The rule involves 'AND'ing the antecedents together and the consequent (class value) for the classification or regression. In this case, the result is the distribution of the available classes (or mean for a numeric value) in the dataset. If this rule does not enclose the test instance, then the default class distributions/value of data that is not enclosed by the rule in the training data is used to predict it. An antecedent is selected by this learner by calculating the Information Gain of each antecedent and the generated rule is pruned using Reduced Error Pruning (REP) or simple pre-pruning depending on the number of antecedents. The weighted mean of the entropies of both the data covered and not covered by the rule is the Information of one antecedent used for classification.

Single conjunctive rule learner [5] is one of the machine learning algorithms and is normally known as inductive Learning. The goal of rule induction is generally to prompt a set of rules from data that captures all general knowledge within that data, and at the same time being as minor as possible Rules can be of various normal forms, and are typically ordered; with ordered rules, the first rule that fires determines the classification outcome and halts the classification process.

Conjunctive rule is an easily inferable 2-class classifier [4]:

$$r_y(x) = \bigwedge_{j \in J} [f_j(x) \lesseqgtr_j \theta_j]$$

Where $f_j(x)$ refers to the features; j {1, . . . ,n} is a features' subset, not very big, usually |j| ≤7; θ_j is threshold; ⋚j is either of the signs ≤ or ≥; y is the rule's class.

If r_y (x) = 1 then 'x' is classified to the class 'y' by the rule 'r'. All objects x are not classified by $r_y$ifr_y (x) = 0.

III. BIG FIVE MODEL

The "Big Five" model for [2] personality prediction is regarded as one of the most researched and well-regarded measures in recent years. The domains of personality of the Big Five model, i.e., Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism, were perceived by Tupes and Christal[7] as the central qualities that appeared from evaluations of earlier personality tests. Extensive study has resulted too many psychologists to admit the Big Five Model as the modern definitive personality model. However, the Big Five model dependency on trait terms directs that the traits are based on a lexical approach to measure the personality.

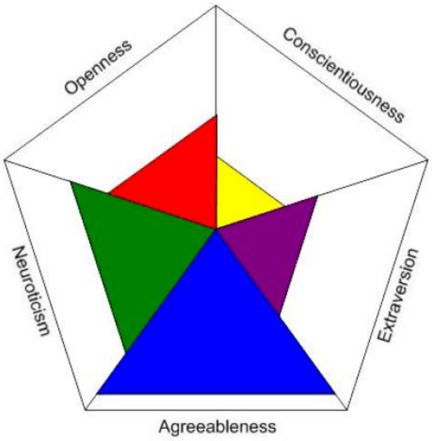

Fig. 1. A person has scores for each of the five personality factors. Together, the five factors represent an individual's personality.

The Big Five traits are characterized by the following:

- Openness: It covers features or dimensions, including active imagination, aesthetic sensitivity, attentiveness to inner feelings, preference for variety, and intellectual curiosity. High scorers tend to be artistic and sophisticated in taste and appreciate diverse views, ideas, and experiences.
- Conscientiousness: responsible, organized, persevering. Conscientious individuals are extremely reliable and tend to be high achievers, hard workers, and planners.
- Extroversion: outgoing, amicable, assertive. Friendly and energetic, extroverts draw inspiration from social situations.
- Agreeableness: cooperative, helpful, nurturing. People who score high in agreeableness are peace-keepers who are generally optimistic and trusting of others.
- Neuroticism: anxious, insecure, sensitive. Neurotics are moody, tense, and easily tipped into experiencing negative emotions.

Since personality relates to our lives and the choices we make, extensive study has been done in this domain. This study has resulted in the identification of many relationships. The personality type of an individual is associated with the users one chooses to friend on Facebook. Similarly, personality features can also be used to estimate aspects of relationships, including choice of the partner, attachment level and success. In case of relational conflict, the Big Five traits are associated with surviving responses, vindictiveness, and contemplation. Within the context of advertising and marketing, personality also relates to preferences. Also, in Human-Computer Interaction (HCI), a revolutionary study on the association between personality and interface has been presented. This work was further evolved into ideas of Graphical User Interface (GUI) design. As a result of this, different Graphical User Interfaces were developed to identify introvert and extrovert personality types.

The usefulness of personality profiles within the social as well as professional context has been demonstrated through various studies. There exists a vast range of real world allegations can be made using insights from personality prediction of the global audience.

IV. DATA SET

In this research, the data set used is a Big Five Model personality prediction dataset which was taken from [9]. The dataset consists of answers to a Big Five Personality Test compiled using the Big-Five Factor Markers from the Item Pool of International Personality [8]. The Big Five test of personality consists of fifty statements. Each statement has been rated by thousands of international test takers on how much they agree to a given statement on a five point scale: (1) Disagree (DA), (2) Slightly Disagree (SDA), (3) neutral (N), (4) slightly agree (SA), and (5) agree (A). The characteristics of the data set are summarized in the Table 3. The aim is to draw predictions about a test giver's personality based on previous answer database.

The attributes of the dataset consists of the following:

- The data set contains 19719 observations spanned across 10 responses each from 5 personality trait categories with no missing values reported.
- 'race' the test taker belongs to
- 'age' of the test taker (age < 13 were not recorded)
- 'engnat' refers to if the test taker's native language is English
- 'gender' of the test taker
- 'hand' refers to the hand they write with usually

TABLE III. DATASET SUMMARIZATION

Instances	19719
Attributes	57
Distinct countries	159
Distinct races	14

V. COMPARISON: PERFORMANCE MEASURES

The performance evaluation of classifiers can be made on basis of certain benchmarks such as Accuracy, Scalability and Interpretability where Accuracy is the ability of the model to correctly predict the class label, Scalability is the ability to construct the model efficiently and Interpretability is the ability of the model to provide the insight.

The assessment of the outcome is made based on the following conditions:

A. *Correlation Coefficient (CC):* Indicates how much true value of interest (θ) and the value estimated using the algorithm (θ^) are related. It gives values between −1 and 1, where 0 is no relation, 1 is very strong, linear relation and −1is an inverse linear relation (i.e. bigger values of θ indicate smaller values of θ^, or vice versa).

B. *Mean Absolute Error (MAE):* MAE is the average prediction error that is gives the average of the difference between expected and actual value for each test case. It is computed as shown:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=0}^{n}\left|\theta^{i} - \theta i\right|$$

C. *Root Mean-Squared Error (RMSE):* RMSE is a measure of differences between values expected by an estimator/algorithm and the values actually observed

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(\theta^{i} - \theta i)^{2}}$$

The mean-squared error is a frequently used measure of success for numeric prediction. However, it can only be used to compare models whose errors are measured in the same units.

D. *Relative Absolute Error (RAE):* Unlike RMSE, RAE can be compared between models with errors measured in the dissimilar units. (Assuming, *θ¯* being a mean value of *θ).*

$$RAE = \frac{\sum_{i=0}^{n}\left|\theta^{i} - \theta i\right|}{\sum_{i=0}^{n}\left|\bar{\theta i} - \theta i\right|}$$

E. *Root Relative Squared Error (RRSE):* Like RAE, the relative squared error too can be evaluated between models whose errors are calculated in the different units.

$$RRSE = \frac{\sum_{i=0}^{n}\left(\theta^{i} - \theta i\right)^{2}}{\sum_{i=0}^{n}\left(\bar{\theta i} - \theta i\right)^{2}}$$

WEKA computes the error measures of RAE and RRSE by normalizing with respect to the performance acquired by predicting the classes' prior probabilities as predicted from the training data with a simple Laplace estimator. The error rates are used for numeric prediction rather than classification.

VI. EXPERIMENTAL RESULTS

The above given error measures are deliberated for each of the two machine learning algorithms for every 50 questions' answer in the dataset. The average of the 10 question data in each of the five personality trait's category – Extroversion (E), Neuroticism (N), Agreeableness (A), Conscientiousness(C) and Openness (O) - is computed for each of the five performance measures to compare against each other.

Analyzing the graphs illustrates clearly that the personality attribute of Openness with a higher correlation coefficient was the easiest to predict and classify for both the classification algorithms. Also, Decision table algorithm was found erroneous to predict the personality traits of Conscientiousness and Agreeableness, and also Neuroticism to some extent. While Agreeableness trait gave greater MAE, RME and RRSE values, Conscientiousness trait produced greater RAE value with Agreeableness closely high as well.

TABLE IV. DECISION TABLE CLASSIFIER RESULTS

	E	N	A	C	O
CC	0.6499	0.69545	0.518	0.5155	0.6879
MAE	0.7523	0.72645	0.812	0.7104	0.5755
RMSE	0.9642	0.94155	1.0359	0.9012	0.7627
RAE (%)	68.5777	64.5703	79.28925	80.2893	68.5232
RRSE (%)	76.0014	71.8451	85.5377	85.4635	72.5511

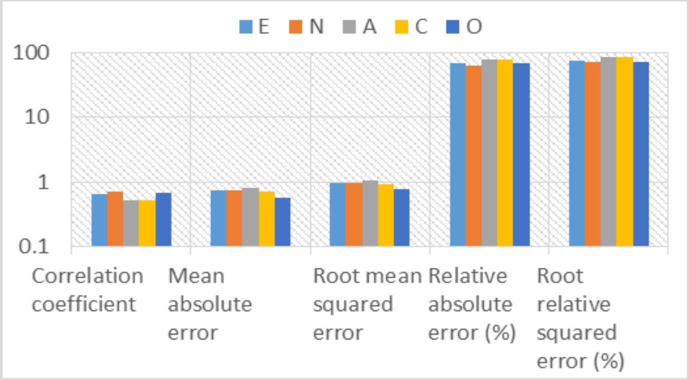

Fig. 2. Comparison Graph for Decision Table classifier

On the other hand, Conjunctive Rule too failed to predict the personality traits of Neuroticism, Agreeableness and Conscientiousness accurately enough. Here, the Neuroticism trait produced a greater MAE and RMSE values while Agreeableness personality trait produced high RAE and RRSE value with Conscientiousness giving closely high error rates.

TABLE V. CONJUNCTIVE RULE CLASSIFIER RESULTS

	E	N	A	C	O
CC	0.4972	0.4752	0.3271	0.3643	0.5432
MAE	0.8867	0.9594	0.9226	0.7578	0.7142
RMSE	1.0994	1.1532	1.1453	0.9827	0.8811
RAE (%)	80.7895	85.2753	89.8446	85.5873	85.5729
RRSE (%)	86.6469	87.9936	94.4634	93.0911	83.9543

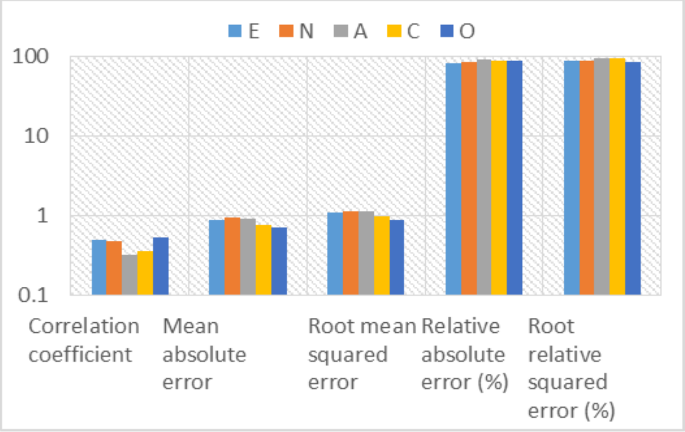

Fig. 3. Comparison Graph for Conjunctive Rule classifier

VII. CONCLUSION

The main aim of this study is to evaluate and investigate two selected classification algorithms and the predictability of personality using WEKA. The Big Five Model Personality data set is used to test the performance of the selected classification algorithms. The algorithm which has the lowest mean absolute error, usually related with a higher accuracy rate, is chosen as the best algorithm. The experimental analysis highlights the fact that though both the algorithms show different accuracy rate for different personality traits in the data set, the personality trait of Openness was the easiest to predict from the Big Five Model. By comparing and contrasting different parameters and the error rate, Decision table clearly outperforms Conjunctive Rule in the rule based classifier algorithm category.

REFERENCES

[1] Kumar, Raj, and Rajesh Verma. "Classification algorithms for data mining: A survey." International Journal of Innovations in Engineering and Technology (IJIET) 1, no. 2 (2012): 7-14.

[2] Golbeck, Jennifer, Cristina Robles, Michon Edmondson, and Karen Turner. "Predicting personality from twitter." In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, pp. 149-156. IEEE, 2011.

[3] Devasena, C. Lakshmi, T. Sumathi, V. V. Gomathi, and M. Hemalatha. "Effectiveness evaluation of rule based classifiers for the classification of iris data set." Bonfring International Journal of Man Machine Interface 1, no. Special Issue Inaugural Special Issue (2011): 05-09.

[4] Konstantin Vorontsov and Andrey Ivahnenko. (2011). “Tight Combinatorial Generalization Bounds for Threshold Conjunction Rules” 4th International Conference on Pattern Recognition and Machine Intelligence (PReMI’11), June 27 – July 1, 2011.

[5] Holmes, Geoffrey, Bernhard Pfahringer, Richard Kirkby, Eibe Frank, and Mark Hall. "Multiclass alternating decision trees." In Machine learning: ECML 2002, pp. 161-172. Springer Berlin Heidelberg, 2002.

[6] Lu, Hongjun, and Hongyan Liu. "Decision tables: Scalable classification exploring RDBMS capabilities." Very Large Data Bases: Proceedings, Cairo, Egypt, IEEE, New York, USA (2000).

[7] E. Tupes and R. Christal. Recurrent personality factors based on traitratings. Journal of Personality, 60(2):225–251, 1992.

[8] Goldberg, Lewis R. "The development of markers for the Big-Five factor structure." Psychological assessment 4.1 (1992): 26.

[9] "Possible Questionnaire Format for Administering the 50-Item Set of IPIP Big-Five Factor Markers". International Personality Item Pool. (http://ipip.ori.org/New_IPIP-50-item-scale.htm)