A Calculation Method of the Similarity Between Trained Model and New Sample by using Gaussian Distribution

Akihiro Matsufuji
Department of Computer Science
Tokyo Metropolitan University
Tokyo, Japan
matsufuji-akihiro@ed.tmu.ac.jp

Haruka Sekino
Department of Computer Science
Tokyo Metropolitan University
Tokyo, Japan
sekino-haruka@ed.tmu.ac.jp

Eri Sato-Shimokawara
Department of Computer Science
Tokyo Metropolitan University
Tokyo, Japan
eri@tmu.ac.jp

Toru Yamaguchi
Department of Computer Science
Tokyo Metropolitan University
Tokyo, Japan
yamachan@tmu.ac.jp

***Abstract*—In many Human Robot Interaction scenarios, social robots are expected to communicate with human naturally. Especially, the skill of predicting human internal state is a field of great interest in many applications, including video surveillance, behavior analysis, human robot interaction, life-logging. While accurately predicting these technologies (e.g. emotion, confident for talking a topic) could have benefits for many fields, generic machine learning systems still yield low performance in some situation. We hypothesize that these sophisticated models suffer from individual differences of human's personality. Therefore, we proposed a multi characteristic model architecture which combines the personalized machine learning models and utilize each model's prediction score in the inference. This architecture formed with reference to ensemble machine learning architecture. In this research, we focus on a similarity between new user and trained user model by using the idea of applicability domain of machine learning models. In the empirical result, we confirmed that data distribution (one way of checking applicability domain) of each user model correspond to the performance of models and we estimated confidence during communication as a human internal state.**

***Index Terms*—affective computing, human robot interaction, non-verbal, multi-modal learning, personal modeling**

I. Introduction

In many Human Robot Interaction (HRI) tasks, social robots are expected to communicate with human naturally compared to the robots for technological supplementary tools for labor intensive or hazardous tasks (e.g., factory automation [1], military operation [2]). They are increasingly designed as social robot to serve as office assistance, teachers, domestic servants, and emotional companions.Social robots are now becoming a part of daily life [3]. Recent social robot could use verbal information to communicate with human (e.g., chatbot [24]) supported by great natural language processing researches. However, our communication use not only verbal information but also non-verbal information and estimating each human's internal state. It is important issue for social robot to communicate with human naturally. Thus, towards building social robots which communicate with human naturally, modeling

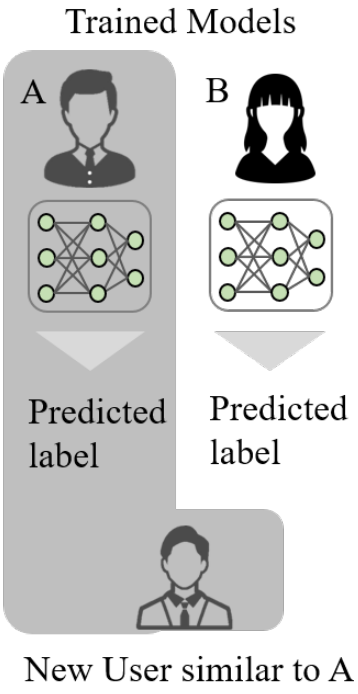

Fig. 1. **Overview of the proposed system.** Our architecture utilized multi-classifier that trained each person's characteristic. We calculated the similarity scores between a new sample and multi characteristics. This similarity score is used as weight scores in ensemble machine learning architecture

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

and predicting human internal state has attracted with many researchers.

Clearly, the skill to predict human internal state in communication (e.g., emotion, talking something confidentially) could be great beneficial [4], [5], especially if such predictions could be made using data collected by using motion and voice information captured by non-contacted sensor. Motion and voice information has attracted attention as important cues for predicting human internal state. According to Merabian [6] from psychology research fields, non-verbal information accounts for 93 % of human message communication. From this, it can be said that non-verbal information plays an important role in human to human communication. Non-verbal information often changes interpretation of the meanings of the language information in the dialog, and it's possible to convey different meanings by changing impressions and voice tones even in the same words. Such a model could open up a range of beneficial application which make predictions about their interesting in communication with human and their understanding in educational robot tutoring system. Unfortunately, modeling human internal state is still an incredibly difficult task and robust system to estimate human internal state has yet to be developed. Historically, classification accuracy is not still enough for all users and environments, even with sophisticated models or multi-modal data. We assume that models suffer from that it is difficult to account for individual differences. Individual differences in personality can strongly affect the human internal state in various scenes. There are various types of human who expression own internal state and what information imply one's human internal state does not apply to everyone else. Practically, a generic machine learning model trained to predict human internal state is inherently limited in the performance it can obtain by using a single model.

Therefore, we proposed the multi characteristic model architecture which utilize various machine learning models trained each type of human's information. Our architecture predicts human internal state by using all multi model's prediction scores. Fig. 1. shows the overview of our multi characteristics model architecture. This model stored several user models and use these models for predicting human internal state adaptively.

In this research, we focus on a calculation of the similarity between new user and trained user model. For building robust predicting human internal state system, the multi characteristics model is necessary to adapt to new user by using appropriate trained models. we consider the applicability domain for estimating relationship between new user and trained user data. We evaluated the calculating applicability domain by comparing with prediction accuracy of each trained model's prediction. The dataset of human internal state, we defined human internal state as whether I am taking confidently or not.

The remainder of this paper is organized as follows: the related works describes in Section 2. Section 3 shows our proposed experimental design, and section 4 explains Results of relationship between prediction accuracy and applicability domain of each models. Conclusions are presented in the last section.

II. Related Works

A. Human Internal State Estimation

The work of Ballihi et al. [8] detects positive/negative emotion from RGB-D data. They classify the intensity of each expression using the upper body motion and face expressions. There is research to estimate the user's comprehension degree by information obtained from facial expressions [9]. According to the experimental results, estimation of the level of users understanding was 70% . Mancini et al. [7] created a real-time system for detecting emotions which were expressed through dancing video in the image processing field. However, the emotional behavior through dancing is largely different from the emotional behavior in daily life. This is based on only information from facial images. However, it is difficult to estimate the degree of comprehension of users with type of poor facial expression. There are studies showing the usefulness of using multi modal information as a method of evaluating the communication skills of people participating in dialog [10]. Therefore, researches on interviewer robots that acquire speech information and motion information and estimate utterance motivation from user's posture and attitude are conducted [11]. From this research, it was found that the individual difference is large for the feature quantity concerning posture. From these studies, it was confirmed that motion information are useful for estimating human internal information.

B. Combining Multi Models for Prediction

For dealing with multi modal information by machine learning model, some researches are separated training sequence part of the architecture for each modal information in a learning model [12]. However, these sophisticated models need large amount of data more than single modal learning model, and it is difficult to consider individual difference. With the regard of utilizing multi machine learning method, the architecture of ensemble machine learning methods [13] are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking). Bagging stands for bootstrap aggregation. One way to reduce the variance of an estimate is to average together multiple estimates [14]. Bagging uses bootstrap sampling to obtain the data subsets for training the base learners. For aggregating the outputs of base learners, bagging uses voting for classification and averaging for regression. In random forests [15], each tree in the ensemble is built from a sample drawn with replacement from the training set. In addition, instead of using all the features, a random subset of features is selected. The bias of the forest increases slightly, but due to the averaging of less correlated trees, its variance decreases, resulting in an overall better model. Boosting [16] referred which a family of algorithms are able to convert weak learners to strong learners. The main principle of boosting is

to fit a sequence of weak learners which models are only slightly better than random guessing to weighted data. In case of the examples were missclassified by earlier rounds, more weight is given to examples. The predictions are then combined through a weighted majority vote (classification) or a weighted sum (regression) to produce the final prediction. The principal difference between boosting and the committee methods, such as bagging, is that base learners are trained in sequence on a weighted data. The advantage of ensemble methods is typically out-performing any machine learning technique. However, ensemble learning has two problems, first, it is difficult to measure correlation between classifiers from different types of machine learning techniques and it is only considered to calculate the weights at training time. so it is difficult to apply to our scenario. Most of ensemble machine learning architectures are modified latter part, integration way to improve the prediction ability. In this paper, we refer the sophisticated integration the multi model part of ensemble learning architecture, we focus on the selection way of several machine learning model for offsetting individual differences.

III. Experiment Design

A. Data Collection

We collected non-verbal information of human internal state at previous work [17]. In this sub-section, the detail of experiment and training data is described. The data for this research were collected by a below experiment. In this research, we set the situation of that participants feel confidence or unconfident as one of human internal states. For this reason, we utilized an agent system to make participants feel confident or unconfident. Fig.2 shows the unconfident situation in communication.

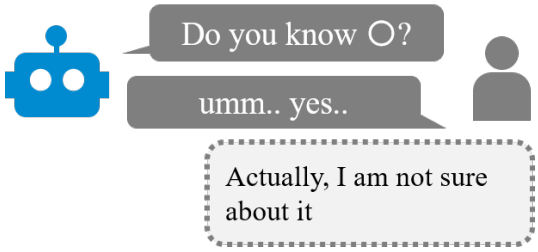

Fig. 2. **Unconfident Situation during Communication**

The experiment involved 11 participants (aged 21-26 years old). Each participant answered 50 questions that we prepare easy version and difficult version. In this study, participants must consider something to answer even if some question are difficult. In this situation, we defined as unconfident situation. All participants agreed to the use of the collected data for the research purpose. Participants answered eight questions one by one, asked by MMD Agent [25], which is a toolkit for building voice interaction systems. To avoid overlap the participants' answer and MMD Agent's utterance, the participants were instructed to answer each question after MMD Agent had terminated the query. Afterward, they filled out the questionnaire from point 1 (Most unconfident) to point 5 (Most confident) to grade their confidence about the answer they gave for each question. This questionnaire is made as Likert-type scale.

The experimental settings are shown in Fig. 3. During the sessions, the behavior of the participants was recorded with a motion tracking camera (Kinect V2, Microsoft). Our system is considered to capture behavior of a participant who seated on a chair. The illumination setting of experimental is similar to general home's. In order to prevent erroneous other person recognition, the experiment settings allow only one participant in motion tracking camera's field of view.

After each session, the participants are asked to answer the question about "which questions did you answer confidently or unconfidently?". We collected questionnaire data as correct labels and motion.

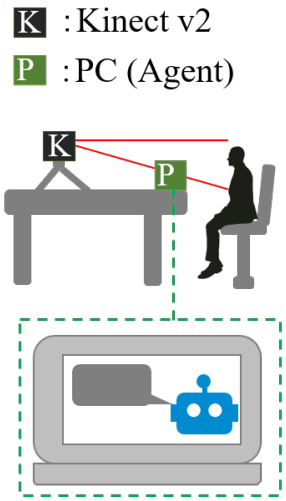

Fig. 3. **Experiment Settings**

B. Preprocessing

In this study, we extracted important cues from obtained motion data for training phase of machine learning architecture. At this subsection, we describe the preprocessing details of motion information. We analyzed the recorded motion as captured 3 dimensional skeleton information. On previous works about motion information, head motion is one of the most prominent social signals either in human-human interaction [19] and human-robot interaction [20], [21]. The consequence of relative researches has proven that the part of head motions shows the most evident results [22]. Thus, we used the movement of the coordinates of the head and gave out an inter frame difference.

C. Training and Implementation

We utilize multilayer perceptron as a machine learning method in this study. We implemented multilayer perceptron algorithm by using machine learning software WEKA [26]. Regarding with the parameters of multilayer perceptron, momentum is 0.2, learning rate is 0.3, epoch is 500. For

evaluating the prediction accuracy of each models, we conduct cross validation method [23]. Our dataset is composed of non-verbal information (motion information) of human internal state, confident or unconfident states. Motion data conclude the head motion about x, y, z axis and cosine degree. We collected above-mentioned data from 11 participants. In training phase, we utilize each participant's features to train each machine learning model, respectively. We prepared the models as many as participants. Each classifier trained only one person's training data which correspond to own model.

D. Prediction of Human Internal State

In this part, classifiers predicted new sample person's internal state from test data of a new sample which trained each person's training data. In our architecture, the number of classifier is same with the number of trained people. In the aspect of variance of ensemble machine learning architecture, individual difference of each trained person is correspond to variance. When our architecture increment machine learning models which trained each person's feature, the variance become larger number. We utilized non-verbal information as training and test data.

E. Similarity Calculation Considering Gaussian Distribution

We aim to calculate the person-person similarity. For building robust predicting human internal state system, the multi characteristics model is necessary to adapt to new user by using appropriate trained models. We considered to utilize the idea of applicability domain [27]for estimating relationship between new user and trained models. Applicability Domain is index of reliability of machine learning models for predicting unknown parameters. More details, it explained the distribution of dataset which trained a model and it tells us which unknown data should not apply for this machine learning models. In our architecture, applicability domain of each models are represents types or individual user that the model are good at.

For the calculating data distribution of the applicability domain, our architecture utilized the Gaussian distribution which is a theoretical distribution with finite mean and standard division. First, the means and standard divisions of each non-verbal features are calculated and generated each feature's Gaussian distributions by using each means and standard divisions. This Gaussian distributions are correspond to each trained models. Second, we prepared the test data for calculating applicability domain by using Gaussian distributions. As the test data, we calculate the first and third quartiles which used in box-whisper plot. Finally, we calculated lower-tail probabilities by using first and third quartiles and Gaussian distributions. In this experimtent, we utilized the lower-tail probabilities by using first and third quarties which calculated by each participant's Gaussian distribution as similarity score between a trained model and new sample.

Before new sample data apply to the Gaussian distribution, these data are normalized to 0-1 scores. Fig. 4 shows about centering.

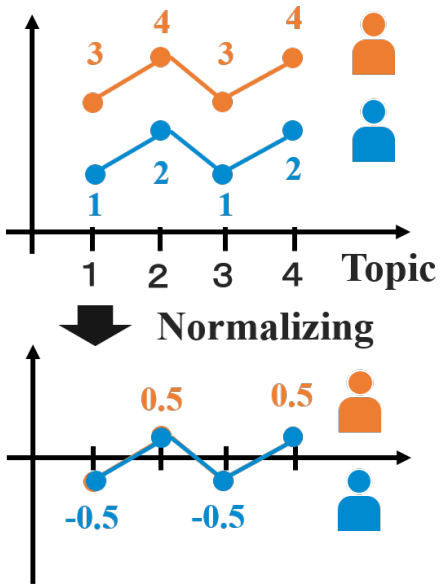

Fig. 4. **Centering the value of each features.**

IV. Result

In this section, we describe about the prediction score of each models and relation between data distribution and prediction result. We evaluated that the calculation method by using data distribution (gaussian distribution) is related to prediction result for each person by using each models.

A. The Prediction Result

We utilize each participant's features to train each machine learning model, respectively. We prepared the models as many as participants. Each classifier trained only one person's training data which correspond to own model. The prediction accuracy of each models shows in table1. This table has 6 out of 11 participants' data. Each participants are described as A, B, C, D, E, F, respectively. The row of the table shows each participants data which utilized as test data. In contrast, the column of the table shows each multilayer perceptron models which trained each person's data. There are several difference between the parameters in the rows and columns of the table, because the meaning of the rows and columns of table 1 is a little bit different as trained models and test subjects. Furthermore, We could clearly divide the "certain" and "uncertain" label from the data in yaw and roll of rotation axis of head by statistical analysis of block whisper plot. Fig 5 shows the participant A's parameter relationship of "certain" and "uncertain" labels. In other person, there are difference of features which divide "certain" and "uncertain" label clearly such as participant A's yaw, roll.

B. The Relation of Data Distribution and Prediction Result

Table 2. shows the probabilities which is the difference between low-tail probabilities of third quartiles and probabilities of first quartiles. This table has also 6 out of 11 participants' data. Each participants are described as A, B, C,

TABLE I
THE PREDICTION ACCURACY OF EACH MODELS

		Test Sample					
		A	B	C	D	E	F
Training Data	A	-	64.0	79.1	65.9	83.3	91.3
	B	37.5	-	60.4	57.4	76.1	80.4
	C	79.2	58.0	-	78.7	80.9	82.6
	D	78.9	68.0	64.6	-	71.4	73.9
	E	56.3	68.0	81.3	46.8	-	76.1
	F	75.0	64.0	79.2	51.1	83.3	-

TABLE II
THE CORRELATION BETWEEN GAUSSIAN DISTRIBUTION AND QUARTILE PARAMETERS

		Third Quartile - First Quartile					
		A	B	C	D	E	F
Gaussian Dist	A	-	1.281	1.684	1.488	1.849	1.762
	B	1.446	-	1.619	1.572	1.823	2.002
	C	1.351	0.966	-	1.755	1.935	1.454
	D	0.723	0.875	1.108	-	1.550	1.173
	E	0.787	0.867	1.234	1.367	-	1.161
	F	1.304	1.147	1.583	1.465	1.223	-

D, E, F, respectively as same with table 1. The row of the table shows each participants data which is probabilities parameter which is the difference between low-tail probabiliries of third quartiles and probabilities of first quartiles. In contrast, the column of the table shows each multilayer perceptron models which trained each person's data.

For evaluating the relation of data distribution and prediction result, we compared between the scores of table 1 and table 2. We checked cols correspond to the participants. For instance, the scores of participant A in table 1 and table 2 of cols are described the relationship with each other participants. Regarding the participants A, the sequence of the number in the table 1 is same with table 2 only except for test data of participants F. In other participants, the prediction accuracy of each models correspond to the relationship between Gaussian distribution and quartile parameters. Regarding with the comparison between data distribution by using Gaussian distribution parameters and prediction result, we checked which test data is good with a trained model. Thus, the columns of the table 2 are selected as indicator of relationship. This parameters apply as the weights of multi characteristics architecture.

However, some of sequence are not correspond to the other one. In this experiment, we calculated simply sum of the relationship by using gaussian distribution. It is further considerable about the method of integrating probabilities.

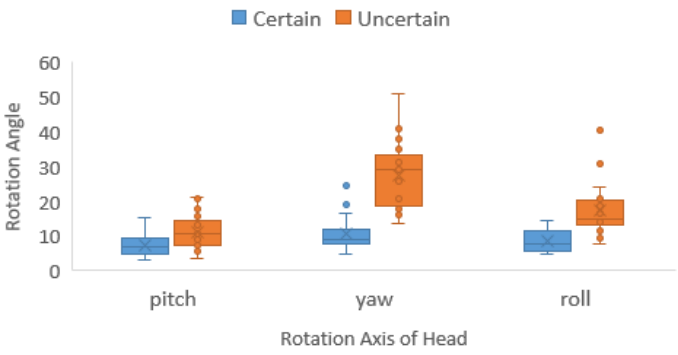

Fig. 5. **The sample of Box Whisper plot.** We could clearly divide the "certain" and "uncertain" label from the data in yaw and roll of rotation axis of head. In other person, there are difference of features which divide "certain" and "uncertain" label clearly such as participant A's yaw, roll.

V. CONCLUSION

We have presented similarity calculation between new user and trained user by using gaussian distribution of each models for a multi model machine learning architecture, that can predict human internal states which affected by personal individual difference. From these insights, we utilize the multi person's characteristic machine learning model which has each human's characteristic type. Our model utilized models which over fitted to each person characteristics. Generally speaking, over fitted model is inappropriate to utilize for classification, because previous works generally considered to create single model to classify in any case. Our approach integrated the any over fitted models. Our architecture estimated human internal state, confident unconfident from motion which is the non-verbal information. In the empirical result, we confirmed that gaussian distribution of each user model correspond to the performance of models.

REFERENCES

[1] A. Cherubini, R. Passama, B. Navarro, M. Sorour, A. Khelloufi, O. Mazhar, P. Fraisse, "A collaborative robot for the factory of the future: BAZAR," The International Journal of Advanced Manufacturing Technology, pp.1–17, 2019.

[2] B. Choi, W. Lee, G. Park, Y. Lee, J. Min, S. Hong, "Development and control of a military rescue robot for casualty extraction task," Journal of Field Robotics, Vol. 36, No. 4, pp. 656–676, 2019.

[3] M. L. Walters, S. Syrdal, K. Dautenhan, R. Boekhorst, K. L. Koay, "Avoiding the uncanny valley, robot appearance personality and consistency of behaivior in an attention seeking home scenario for a robot companion," Autonomous Robots, Vol.24, No.2, pp.159–178, 2008.

[4] E. Marinoiu, M. Zanfir, V. Olaru, G. Sminchisescu, "3d human sensing, action and emotion recognition in robot assisted therapy of children with autism," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition ,pp. 2158–2167, 2018.

[5] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, S. Wermter, "On the Robustness of Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks," In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 854–860, 2018.

[6] Merabian, A, "Silent Communication," Wadsworth, Belmont, Calfornia, 1971.

[7] M. Mancini, G. Castellano, "Real-Time Analysis and Synthesis of Emotional Gesture Expressivity," In Proceedings of the Doctoral Consortium of 2nd International Conference on Affective Computing and Intelligent Interaction.

[8] L. Ballihi, A. Lablack, B. B. Amor, I. M. Biasco, M. Daoudi, "Positive Negative Emotion Detection from RGB-D Upper Body Images," Face and Facial Expression Recognition from Real World Videos", Springer International Publishing, vol. 8912, pp. 109–120, 2015.

[9] A. Mimura, M Hagiwara, "Understanding Presumption System from Facial Images," The transaction of the institute of Electrical Engineers of Japan, Vol.120, No.2, pp.273–278, 2000.

[10] S. Okada, Y. Matsugi, Y. Nakano, Y. Hayashi, H. H. Huang, Y. Takase, K. Nitta, "Estimating Communication Skills based on Multimodal Information in Group Discussions," Journal of Japanese Society for Artificial Intelligence, Vol.31, No.6, pp.AI30-E112, 2016.

[11] F. Nagasawa, T. Ishihara, S. Okada, K. Nitta, “A Case Study Toward Implementing Adaptive Interview Strategy Based on User’s Attitude Recognition for a Interview Robot,” Journal of Japanese Society for Artificial Intelligence, Vol.31, 2H4-OS-35b-1, 2017.
[12] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, “Multimodal deep learning”, In Proceedings of the 28th international conference on machine learning (ICML-11), pp. 689–696, 2011.
[13] TG. Dietter, G. Thomas, “Ensemble methods in machine learning,” In Multiple classifier systems, Vol. 1857, pp.1–15, 2000.
[14] S. Quan P. Bernhard, “Bagging ensemble selection,” AI 2011: Ad vances in Artificial Intelligence, pp.251–260, 2011.
[15] L.Breiman, “Random forests,” In Machine learning, Vo. 45, No. 1, pp.5-32, 2001.
[16] D. G. Thomas, “An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization”, In Machine learning, No. 40, Vol. 2, pp.139–157, 2000.
[17] E. Kasano, S. Muramatsu, A. Matsufuji, E. Sato-Shimokawara, T. Yamaguchi, “Estimation of Speakers Confidence in Conversation Using Speech Information and Head Motion,” the 16th international conference on ubiquitous robots, 2019.
[18] P. Boersma, “A System for Doing Phonetics by Computer,” Glot International, Vol. 34, No. 5, pp. 9–10, 2001.
[19] A. Vinciarelli, M. Pantic, B. Hereve, “Social signal processing: survey of an emerging domain, Image and Vision Computing”, Vol. 27, No. 12, pp. 1743–1759, 2009.
[20] M. Giuliani, N. Mirning, G. Stollnberger, S. Stadler, R. Buchner, M. Tscheligi, “Systematic analysis of video data from different human robot interaction studies: a categorization of social signals during error situations,” Frontiers in psychology, Vol. 6, pp. 931, 2015.
[21] R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, A. Waibel, “Natural human-robot interaction using speech, head pose and gestures,” In Intelligent Robots and Systems.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Confidence Vol. 3, pp. 2422–2427, 2004.
[22] L. P. Morency, C. Sidner, C. Lee, T. Darrel, “Head gestures for perceptual interfaces: The role of context in improving recognition”, Artificial Intelligence, Vol. 171, No. 8–9, pp. 568–585, 2007.
[23] D.M. Allen, “The relationship between variable selection and data augmentation and a method for prediction,” in technometrics, Vol. 16, pp-125–127, 1974.
[24] Meyer von Wolff, Raphael, Sebastian Hobert, and Matthias Schumann. ”How May I Help You?–State of the Art and Open Research Questions for Chatbots at the Digital Workplace.” Proceedings of the 52nd Hawaii International Conference on System Sciences. 2019.
[25] A. Lee, K. Oura, K. Tokuda, MMDAgent - A fully open-source toolkit for voice interaction systems, Proceedings of the ICASSP 2013, pp. 8382-8385, 2013.
[26] E. Frank, MA. Hall, H. Ian, The WEKA Workbench: Online Appendix for ”Data Mining: Practical Machine Learning Tools and Techniques”, Morgan Kaufmann, Fourth Edition, 2016.
[27] L. Ruiz, Irane, and M.Á. Gómez-Nieto, Building of robust and interpretable QSAR classification models by means of the rivality index, Journal of chemical information and modeling Vol. 59, No. 6, pp. 2785-2804, 2019.