Paper:

Dynamically Weighted Ensemble Models Based on the Behavioral Similarity Towards Sentiment Estimation

Akihiro Matsufuji[*1], **Eri Sato-Shimokawara**[*1], **Toru Yamaguchi**[*1], **Lieu-Hen Chen**[*2]

Tokyo Metropolitan University, Tokyo Japan[*1]
National Chi Nan University, Nantou Taiwan[*2]
E-mail: matsufuji-akihiro@ed.tmu.ac.jp, [eri, yamachan]@tmu.ac.jp, lhchen@csie.ncnu.edu.tw
[Received 00/00/00; accepted 00/00/00]

Abstract. Recent emotion recognition applications strongly rely on supervised learning techniques for the distinction of general emotion expressions. However, they are not reliable to unknown person, due to the individual differences between human emotional state and their emotion expressions. In this study, a novel multi classifier ensemble learning using dynamic weights based on the similarity of the emotion expression among different people is proposed to reduce the interference of unreliable decision information and adapt their individual differences. We proposed two dynamic weights definition methods which are based on the statistically feature analysis and the analysis of multi modal time series data using image-conversion. We demonstrated the flexibility of these methods to adapt the their individual differences with finding the trained data of person who has similar individual differences between human emotional state and their expressed features.

Keywords: Affective computing, Multi classifier ensemble, Dynamic weights, Individual difference, Visualization

1. Introduction

Research on the analysis and modeling of human emotional behavior are known as affective computing. Emotions play an important role in all aspects of human life, especially, our work performance, decision making, mental health [1]; the applications of human-robot interaction with accurate emotion recognition could change the dialog strategy corresponding to human emotion state.

However, the performance of the machine learning models for emotion recognition has recently stagnated [2]. The most major cause is the individual differences between human emotion state and their expressed features. The diversity of emotion expressions becomes even more complex when we take into consideration inter-personal characteristics (e.g., personality, mood, genetics and cultural background) [4][5]. Thus, it is very hard to adapt end-to-end learning-based models to different individual's emotion expression, mostly due to very costly training process, especially, data collection in a unobtrusive and privacy-sensitive way. In the real-world application, these models need to learn a new relationship between human emotion state and their expressed features, even of sophisticated classifiers need to be re-trained.

To solve the personalization of emotion expression problems, unsupervised dynamic adaptation were explored [7]. A classifier was trained with data from an individual person to improve emotional representation. However, these models are not able to adapt to unknown person to online and continuous learning scenarios found on real-world applications.

Therefore, we address the problem of learning adaptable individual's emotion expressions by focusing on improving emotion recognition based on the report that they can adapt their own perception to how that specific person expresses emotions when humans already know a specific person [6]. Thus, we assumed that humans tend to rely on their own prior knowledge of other people's expressions learned over to understand individual's expressions at the beginning of a dialogue. We propose the use of a ensemble model including multi classifier trained with data from each individual person, and define the weights of classifiers dynamically based on the similarity of personalized expressions between the trained person and unknown person. Our contribution is that we proposed the two dynamic weights definition methods based on the calculating the similarity of personalized expressions.

2. Related Work

2.1. Supervised Learning of Diversity

Some researches focused on the problem of learning diversity on emotion expressions was with the database including emotion expression in the wild (e.g. AffectNet [8], Emotiw18 [9]). These research use a large amount of data to increase the variability of emotion representations. Although deep learning models trained with these dataset improved the performance [10], they still suffer from the lack of adaptability to personalized expressions. Koldijk et al. [11] presented personalized stress estimator using participants' ID as one of explanatory variable. Canzian et al. [12] found that the performance of classifier trained with data from specific person is over than the performance of the classifier trained with data from various people on the mood estimation task. These researched showed the importance of considering personalized emotion expressions, but, it was needed to train the target person's expressions or ID.

2.2. Multi Classifier Ensemble

Ensemble learning methods usually combine multiple base classifiers, which are generated by varying the training sample, the parameters, and achieve a forecasting result with higher stability and accuracy [13][14]. To improve the performance of the ensemble learning methods, many efforts have been made to build a more effective static combiner which could be separated into fixed method (e.g., mean, majority voting), trainable method (e.g., fisher linear discriminant). However, they combine the classifiers for the testing samples by using the same combine rule, and it could not be applied to adapt the difference between testing targets. On the other hand, Zhu et al. [15] proposed the dynamic weighting ensemble based on the cross-validation, which assign different weights to the base classifiers for different test samples. However, this method does not consider the law of change of base classifiers' classification ability and it still weights the predictions of classifiers based on their past performance. Thus, it is difficult to be applied to adapt the personalized expressions. For dealing with concept drift, Fan et al. [16] prepared the base classifiers with the training data separating by obtained time, and dynamically assign weights to the base classifiers for each test samples based on the time when the test data were obtained. Inspired by the above weights definition method based on time, we proposed the weights definition method using the similarity of individuals' personalized expressions.

3. Proposed Method

3.1. Multi Person Ensemble Learning Model

In this sub-section, the procedure of the constructing a ensemble model including multi classifier trained with data from each individual person and these definitions are described. The following procedure is introduced as a binary classification task. Given the data from people denoted as $1, 2, ..., p \in P$, the data of person A is denoted as $D_p = [(X_{1,1}, X_{1,2}, ..., X_{1,f}, y_1), ..., (X_{t,1}, X_{t,2}, ..., X_{t,f}, y_i)]$, where $[X_{t,f}, y_t]$ is a data instance, $y_T \in [1, -1]$ is a binary label data, and f and t represent the number of expressed features and time series data described as frames, respectively. A set of classifiers denoted as $E = [C_1, C_2, ..., C_p]$, where E is an ensemble learning model and C_p is a classifier trained with data from person P, and $Y = [y^1_{pred}, y^2_{pred}, ..., y^p_{pred}]$ and $W = [w^1, w^2, ..., w^p]$ are the classification results and estimated weight corresponding to p classifiers. The final prediction of our ensemble learning model y_{pred} is calculated by using Y and W shown as follows:

$$y^p_{pred} = C_p(x_{test}) \quad \text{. . . . .} \quad (1)$$

$$y_{pred} = \frac{1}{P}\sum_{p=0}^{P} y^p_{pred} \times w^p \quad \text{. . . . .} \quad (2)$$

In the following sub-sections, the two methods of weights calculation are described. The method 1 used the statistics values of time-series data used as training data of machine learning for calculating the weights and it is assumed to be high affinity to the performance of classifiers. On the other hand, the method 2 used the raw data of multi modal time series data for detailed analysis of the behavior to calculate the similarity weights.

3.2. Method 1: Person-Feature Importance Matrix

As for calculating the similarity between people, collaborative filtering has become one of the most used approaches to provide personalized services for users such as an item recommendation [17]. The key of this approach is to find similar people using person-item rating matrix and similarity algorithms (e.g., cosine and Pearson correlation coefficient). From these insights, we proposed a person-feature importance matrix to find person who has similar personalized feature expressions with cosine similarity algorithms. Note that this method require the feature represented as statistics values such as mean, standard deviation, maximum and minimum values of time series data which widely are used as features in the machine learning tasks [19].

To create a person-feature importance matrix from features represented as statistics values, we calculating the mutual information which has been widely used to find the importance feature to contribute to the predicting labels in machine learning [18]. Given the data from person p, $X^p_f = [x_1, x_2, ..., x_e]$ where x_e is a statistics values in an exchange in a dialogue and e and f are the number of exchange and features. $Y^p = [y_1, y_2, ..., y_e]$ are the label data. The mutual information of features X^p_f and label Y^p is denoted as:

$$MI(X^p_f; Y^p) = \sum_{i=0}^{e}\sum_{j=0}^{e} P(x_i, y_i) \log \frac{P(x_i|y_i)}{P(x_i)} \quad \text{. .} \quad (3)$$

A person vector in a person-feature importance matrix is represented as mutual information of all feature. Thus, this person vector described the personalized expression based on the calculating the correlation between features and emotional state.

$$\vec{x^p} = [MI(X^p_0; Y^p), MI(X^p_1; Y^p), ..., MI(X^p_f; Y^p)] \quad (4)$$

Given the few test samples to adapt the test target, The similarity of personalized expressions could be calculated using cosine similarity algorithm.

$$w^i = cos(\vec{x^p}, \vec{x^{test}}) = \frac{\vec{x^p} \cdot \vec{x^{test}}}{|\vec{x^p}||\vec{x^{test}}|} \quad \text{. . . . . .} \quad (5)$$

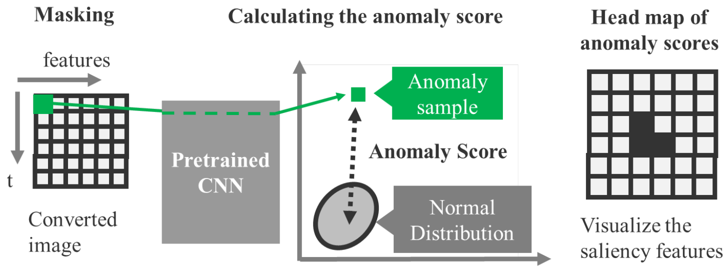

Fig. 1. : Image-conversion of Multi Modal Time Series data

3.3. Method 2: Image-Conversion based Multi Modal Time Series Data Analysis

While the features represented as statistics values of time series data widely are used in machine learning, the interactions among features in the time series data is also important factor to describe the characteristics of personalized emotion expressions. Idealy, the calculation of similarity among people should be extended to the multi modal time series data analysis. Therefore, we proposed a image-conversion based multi modal time series data analysis. The vertical axis of an converted image from multi modal time series data is a frame/time information and the horizontal axis of an image sets the multi modal features. Furthermore, we take advantage of the image processing techniques of explanation to find the saliency features from an image represents multi modal time series information described as Fig. 1.

Some explanation method of object detection task systematically occluded different portions of the input image with a grey square, and monitor the change of the class probability of the trained classifier [20][21]. When the class probability decreased, the occluded portions of the input image is importance area to predict the specific label. However, these methods required the trained model which outputs the class probability; it is difficult to apply to the adaptation using small amount of data. Thus, we used the anomaly detection framework to calculate the mahalanobis distance between binary labels by modeling the distribution of a label [22]. Then, we monitor the change of the output distance between image features belong to binary labels when systematically occluded different portion of the image with a gray square. Similarity to the previous work, the output distance between binary labels decreased, the occluded portions of the input image is a saliency feature of multi modal time series data. We calculated the similarity of the changing the output among people using dynamic time warping method which calculate the distance between time series information [23].

4. Experiments

4.1. Data

We collected non-verbal data of the internal state of humans in a previous study [24].Ten participants (aged 21-26 year) were recruited from the Tokyo Metropolitan University. Each participant answered 50 questions from several fields (e.g., history and the seasons) asked by an agent[25]. Afterward, they filled out a questionnaire to annotate their confidence of the answer. This questionnaire was created on a 5-point Likert scale.

4.1.1. Motion Features

The time series motion data of the head recorded by the Microsoft Kinect sensor, and the data were normalized for each participant through the Z score normalization, that is, considering a mean and standard deviation of zero and one, respectively, for all samples pertaining to each participant. The mean, standard variation, maximum values of calculated velocity and acceleration were used as the motion features to train the machine learning model. The difference between the frames are calculated from the recorded time series data for image conversion method. All the features were extracted from the whole dialogue per exchange.

4.1.2. Annotation

The participants themselves annotated the labels per exchange on a 5-point Likert scale. We used only 1 ("I did not have the confidence") points and 5 ("I had the confidence") points from the questionnaire as binary classification labels, and we perform the synthetic minority oversampling technique (SMOTE) [26] to handle the class-imbalanced data.

4.2. Evaluation Procedure

To evaluate the models, the leave one person out cross-validation (LOPOCV) was performed in the logistic regression classifiers . In the LOPOCV, the samples corresponding to each exchange between the person and dialogue system were used as the test data, and the remaining samples were used as the training data. This procedure ensured that the test data from one person were completely excluded in the training dataset, thereby avoiding overestimation. The baseline is a classifier trained with training dataset, proposed methods used the base classifiers trained with data from each person and combine these classifiers' result by using each weight definition method.

5. Result

The Table 1 shows the evaluation results of LOPOCV, and the average accuracy of both proposed dynamic weights ensemble models are higher than the baseline model. Whereas, as for the accuracy with each validation data, there were a validaton data with low accuracy in all models. It can be considered these validation data was inherently difficult to classify only with used features in our experiment. Furthermore, the some validation data with low accuracy was not founded the person who has similar personalized expressions with validation appropriately due to the small number of the person data in this study.

Table 1. : Comparison of prediction accuracy

Validation data	Baseline	Proposed 1	Proposed 2
Person A	0.750	0.833	0.750
Person B	0.818	0.909	0.909
Person C	0.545	0.545	0.636
Person D	0.571	0.500	0.571
Person E	0.417	0.500	0.583
Person F	0.786	0.857	0.857
Person G	0.428	0.571	0.571
Person H	0.400	0.400	0.700
Person I	0.727	0.727	0.727
Person J	0.788	0.788	0.788
Average	0.623	0.670	0.700

6. Conclusion

We presented a multi classifier ensemble learning using two kind of the dynamic weights based on the similarity of the personalized expression among different people. That are a person-feature importance matrix which vectorized the importance of the features represented as statistics, and an image-conversion based multi modal time series which take advantage of the image processing explanation methods. We demonstrated the flexibility of these proposed methods to adapt the test data from unknown person with finding the trained data of person who has similar personalized expressions using few samples from the person. In the future work, we will examine the data from participants of wider age distribution.

References:

[1] S. L. Koole, K. Rothermund, “i feel better but i don’t know why,” The psychology of implicit emotion regulation. Cognition and Emotion, Vol.25, No.3, pp.389–399, 2011.

[2] S. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. W. Picard, Personalized multitask learning for predicting tomorrows mood, stress, and health, IEEE Trans. on Affective Computing, Vol. 11, No. 2, pp. 200-213, 2020

[3] J. A. Russell, Cross-cultural similarities and differences in affective processing and expression. In Emotions and Affect in Human Factors and Human-Computer Interaction, pp. 123–141, Elsevier, 2017.

[4] M. Gendron, D. Roberson, J. M. van der Vyver, and L. F. Barrett, Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture, Emotion, Vol. 14, No. 2, pp. 251-262, 2004.

[5] J. A. Russell, Cross-cultural similarities and differences in affective processing and expression. In Emotions and Affect in Human Factors and Human-Computer Interaction, pp. 123–141, Elsevier, 2017.

[6] S. Hamann, and T. Canli, Individual differences in emotion processing. Current opinion in neurobiology, Vol. 14, No.2, pp.233–238, 2004.

[7] Y.A. Chen, J.C. Wang, Y.H. Yang, and H. Chen, Linear regression-based adaptation of music emotion recognition models for personalization. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 2149–2153. IEEE, 2014.

[8] A. Mollahosseini, B. Hasani, and M. H. Mahoor, Affectnet: A database for facial expression, valence, and arousal computing in the wild. arXiv preprint arXiv:1708.03985, 2017.

[9] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In Proceedings of the 2018 on International Conference on Multimodal Interaction, pp. 653– 656. ACM, 2018.

[10] D. Kollias, and S. Zafeiriou, Training deep neural networks with different datasets in-the-wild: The emotion recognition paradigm. In 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE, 2018.

[11] S. Koldijk, M. A. Neerincx, and W. Kraaij, Detecting work stress in offices by combining unobtrusive sensors, In Journal of IEEE Transactions on Affective Computing, Vol. 9, No. 2, pp.227–239, 2017.

[12] L. Canzian, and M. Musolesi, Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis, In Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing, pp. 1293-1304, 2015.

[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: Advances in Artificial Intelligence, pp.251–260, 2011.

[15] Y. Zhu, J. Ou, G. Chen, H. Yu, Dynamic weighting ensemble classifiers based on cross-validation. Neural ComputAppl, Vol.20, No. 3, pp.309–317, 2011.

[16] D. Fan, L. Jie, G. Zhang, K. Li, Active fuzzy weighting ensemble for dealing with concept drift International Journal of Computational. Intelligent. Systems, Vol.11, pp.438-450, 2011.

[17] M. Schwarz, M. Lobur, and Y. Stekh, Analysis of the effectiveness of similarity measures for recommender systems. In 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), pp. 275-277, 2017.

[18] H. Peng, L. Fuhui, and D. Chris, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, Vol. 27, No. 8, pp. 1226-1238, 2005.

[19] S. Katada, S. Okada, Y. Hirano, K. Komatani, Is She Truly Enjoying the Conversation?: Analysis of Physiological Signals toward Adaptive Dialogue Systems. International Conference on Multimodal Interaction (ICMI), pp.315-323, 2020.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, ”Why should i trust you?” Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144, 2016.

[21] M. D. Zeiler, and R. Fergus, Visualizing and understanding convolutional networks. In European conference on computer vision, pp. 818-833, 2014.

[22] O. Rippel, P. Mertens, and D. Merhof, Modeling the distribution of normal data in pre-trained deep features for anomaly detection. In 2020 25th International Conference on Pattern Recognition (ICPR), pp. 6726-6733, 2021.

[23] D. J. Berndt, and J. Clifford, Using dynamic time warping to find patterns in time series. In KDD workshop, Vol.10, No.16, pp. 359-370, 1994.

[24] 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.

[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] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, Vol.16, pp.321-357, 2002.