An End-to-End Traditional Chinese Medicine Constitution Assessment System Based on Multimodal Clinical Feature Representation and Fusion

Huisheng Mao[1*], **Baozheng Zhang**[1,2*], **Hua Xu** [1* †], **Kai Gao** [2]

[1] State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

[2] School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China

mhs20@mails.tsinghua.edu.cn, zhangbaozheng@tom.com, xuhua@tsinghua.edu.cn, gaokai@hebust.edu.cn

Abstract

Traditional Chinese Medicine (TCM) constitution is a fundamental concept in TCM theory. It is determined by multimodal TCM clinical features which, in turn, are obtained from TCM clinical information of image (face, tongue, etc.), audio (pulse and voice), and text (inquiry) modality. The auto assessment of TCM constitution is faced with two major challenges: (1) learning discriminative TCM clinical feature representations; (2) jointly processing the features using multimodal fusion techniques. The TCM Constitution Assessment System (TCM-CAS) is proposed to provide an end-to-end solution to this task, along with auxiliary functions to aid TCM researchers. To improve the results of TCM constitution prediction, the system combines multiple machine learning algorithms such as facial landmark detection, image segmentation, graph neural networks and multimodal fusion. Extensive experiments are conducted on a four-category multimodal TCM constitution dataset, and the proposed method achieves state-of-the-art accuracy. Provided with datasets containing annotations of diseases, the system can also perform automatic disease diagnosis from a TCM perspective.

Introduction

Traditional Chinese Medicine (TCM) Constitution (TCMC) is a fundamental concept in TCM theory. It lays the foundation for disease diagnosis and treatment. In order to judge a patient's TCMC, multiple sets of TCM clinical features, such as the patient's physique, complexion, tongue color, and emotional state, need to be considered. The features are usually collected by experienced TCM doctors, using the 4 diagnostic methods of TCM: inspection, listening and smelling examination, inquiry, and pulse taking and palpation (Hu and Liu 2012). To provide an end-to-end solution to TCMC assessment task, such features need to be learned by an algorithm from TCM clinical information of image, audio and text modality. What's more, the learned multimodal representations need to be comprehensively processed with multimodal fusion methods. These are two main challenges in end-to-end TCMC assessment.

The TCM Constitution Assessment System (TCM-CAS) is an end-to-end system that can automatically assess a patient's TCMC by processing TCM clinical information collected on-site. Due to the restrictions of training data, the system can only assess 4 out of the 9 TCMCs at the moment. Also, the input is reduced to two modalities, where a facial image, a tongue-dorsal image and a questionnaire are required, as shown in Figure 1. Given these inputs, the system will generate a report where the final prediction of TCMC along with some intermediate results are presented, as shown in Figure 2. With enough data, the system can be easily extended to assess all nine TCMCs as well as other TCM classification tasks such as disease diagnosis. The system also features auxiliary functions such as TCM data exploring and TCM clinical feature analysis to better aid TCM researchers. The system is fully demonstrated in this video[1] and the source code is available at Github[2].

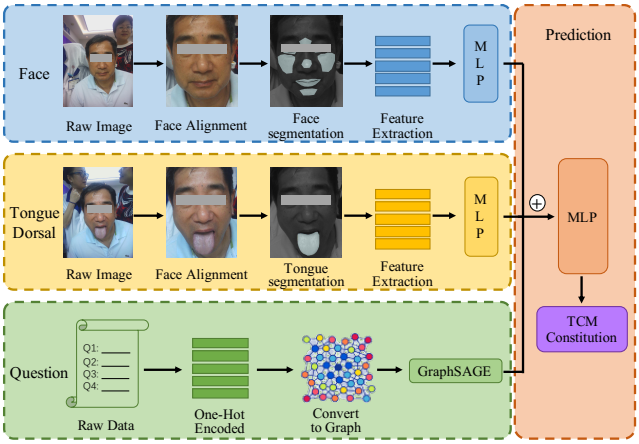

Figure 1: The overall structure of the end-to-end TCM constitution assessment algorithm used in the system.

Method

As shown in Figure 1, the TCMC assessment algorithm consists of four modules: face, tongue-dorsal, question, and prediction. The first three modules process input TCM clinical information and learn discriminative feature representations,

* These authors contributed equally to this work.

† Hua Xu is the corresponding author.

[1] Video: https://youtu.be/n19R_D21X2Q

[2] Source code: https://github.com/thuiar/TCM-CAS

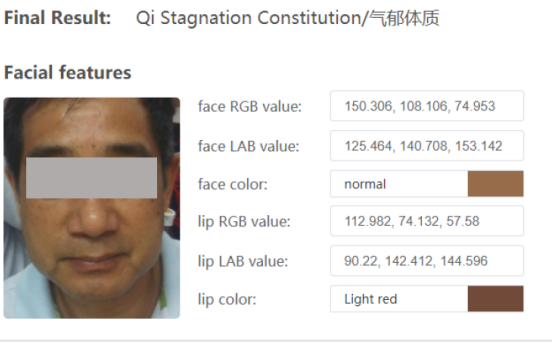

Figure 2: Part of the TCMC report generated by the system.

while the last one jointly processes the learned multimodal features and predict TCMC.

Face Module

The face module requires an image as input. The largest human face that appears in the image is detected, aligned, and segmented. These steps are done based on the facial landmark detection algorithm provided by Mediapipe (Lugaresi et al. 2019). The segmented areas are then passed to a multilayer perceptron (MLP) to learn the representations. Two facial TCM clinical features, namely facial color and lip color, are extracted from the representations. The features will be presented as intermediate results for better interpretability.

Tongue-Dorsal Module

The tongue-dorsal module deals with the input image in a similar approach to the face module, except that the tongue-segmentation is based on MiniSeg model (Qiu et al. 2021). Tongue-dorsal TCM clinical features extracted by this module include tongue-coating color and tongue color.

Question Module

The question module learns feature representations from input data of question-answer pairs. The 15 single or multiple-choice questions are meticulously designed by experts, for example, whether the patient is prone to fatigue and the sleeping quality of the patient.

Inspired by Yao, Mao, and Luo (2019), we build a heterogeneous graph that contains patient and symptom nodes. As shown in Figure 3, the existence of a patient-symptom edge shows that the patient has the symptom. A symptom-symptom edge indicates co-occurrence of the two symptoms. The weight on a symptom-symptom edge pointing from A to B shows the probability of symptom B given symptom A on a patient. The graph is initialized with one-hot representations for patient and symptom nodes. A GraphSAGE (Hamilton, Ying, and Leskovec 2017) model is used to learn the features for patient nodes, as supervised by TCMC labels.

Prediction

After representations are learned in the previous modules, the prediction module concatenates the features and passes them through an MLP to perform TCMC classification. Other multimodal fusion methods, such as Tensor Fusion Network (Zadeh et al. 2017) and Low-rank Multimodal Fusion (Liu et al. 2018) are also experimented upon. However, the results are not satisfactory compared to the simple fusion method of concatenation, as shown in Figure1.

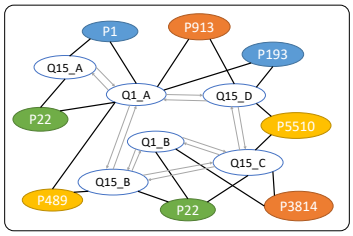

Figure 3: Schematic of the patient-symptom graph. Nodes start with "P" are patient nodes, the color indicates the TCMC of the patient. Nodes start with "Q" are symptom nodes, each corresponding to a choice of a question. The bold black edges are bi-directed patient-symptom edges, while thin gray edges are symptom-symptom edges.

Model	Acc	F1
MLP-unimodal	86.12	86.01
GraphSAGE-unimodal	86.38	86.16
XGBoost-multimodal	89.60	89.48
MLP-multimodal	90.15	90.02
SE-Resnet-multimodal	90.37	90.21
GraphSAGE-multimodal-TFN	91.56	91.48
GraphSAGE-multimodal-concat	**92.07**	**91.96**

Table 1: 5-seed average results on the TCMC dataset. "unimodal" refers to the text modality (question module) alone.

Experiments

To achieve better performance, extensive experiments are conducted on a four-category TCMC dataset which contains 5515 samples. Due to the page limit, only some of the results are listed in table 1. To validate the effectiveness of our GNN-based network, MLP and XGBoost are used to replace GNNs in the question module. In other experiments, our feature extraction methods for the image modality are replaced by different computer vision models. It can be concluded from the results that our method outperforms other models in this specific task.

Conclusion

We present an end-to-end Traditional Chinese Medicine (TCM) Constitution Assessment System. It combines multiple machine learning techniques to process multimodal inputs, learn TCM clinical features and assess patients' TCM constitution. The method used in the system achieves state-of-the-art performance on multimodal TCM constitution classification task. The system can be easily extended to a TCM disease diagnosis system in the future.

Acknowledgements

This paper is founded by National Key R&D Program Projects of China (Grant No: 2018YFC1707605).

References

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 1025–1035.

Hu, J.; and Liu, B. 2012. The basic theory, diagnostic, and therapeutic system of traditional Chinese medicine and the challenges they bring to statistics. *Statistics in medicine*, 31(7): 602–605.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A. B.; and Morency, L.-P. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2247–2256.

Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M. G.; Lee, J.; et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Qiu, Y.; Liu, Y.; Li, S.; and Xu, J. 2021. MiniSeg: An Extremely Minimum Network for Efficient COVID-19 Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4846–4854.

Yao, L.; Mao, C.; and Luo, Y. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 7370–7377.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114.