Voting Ensemble Pruning for De-identification of Electronic Health Records

MUSC
MEDICAL UNIVERSITY
of SOUTH CAROLINA

Youngjun Kim[1], PhD, Stéphane M. Meystre, MD, PhD[1,2]

[1] Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC
[2] Clinacuity, Inc., Charleston, SC

Abstract

Ensemble pruning aims to reduce ensemble members for better performance and save computational costs.

Our pruning method can automatically determine the voting threshold and the optimal combination of information extraction models.

In the iterative pruning process, performance degradation is prevented by excluding a model that is least helpful at each step.

The application for text de-identification showed that the **pruned voting ensemble achieved a higher performance** than an ensemble using all models.

Introduction

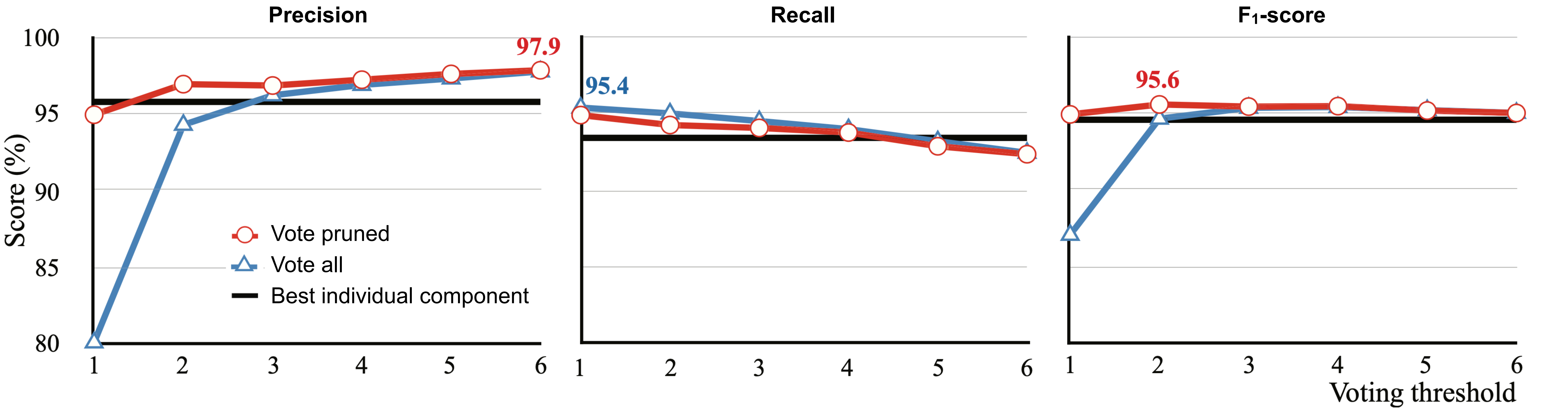
Ensemble pruning aims to reduce ensemble members to achieve similar or better performance than the ensemble of all members. It can offer the advantage of saving computational time or physical resources that would be consumed by excluded members.

In our previous research on electronic health record (EHR) narratives de-identification[1], we applied a variety of information extraction methods including deep learning, shallow learning, and rule-based approaches.

We created a voting ensemble method that combines **twelve de-identification models**.

Although the voting ensemble yielded better performance than individual models, ablation tests revealed that some models did not contribute to the performance of the voting ensemble.

In this study, we present a **pruning method that allows one to automatically determine** the voting threshold (i.e., number of members voting for one annotation) and **the optimal combination of de-identification models**.



Methods

The input for this method is the predictions from each model.

We performed 10-fold cross-validation on the 2014 i2b2 NLP challenge[2] training set to acquire these predictions. This procedure is to maximize the micro-averaged F_1 score with strict entity matching, where both the text span and identifiers category exactly match the reference annotations.

- We began with an ensemble containing all individual models ranked with the F_1 score measured during cross-validation on the training set.
- At each step, we subsequently excluded one model such that the F_1 score was the highest without the model.
- These steps continued until the performance was no longer improved. This procedure was repeated for each voting threshold to obtain a subset of de-identification models.

We chose the voting threshold that yielded the best performance.

Acknowledgments: Research supported by the NIGMS (R41GM116479).

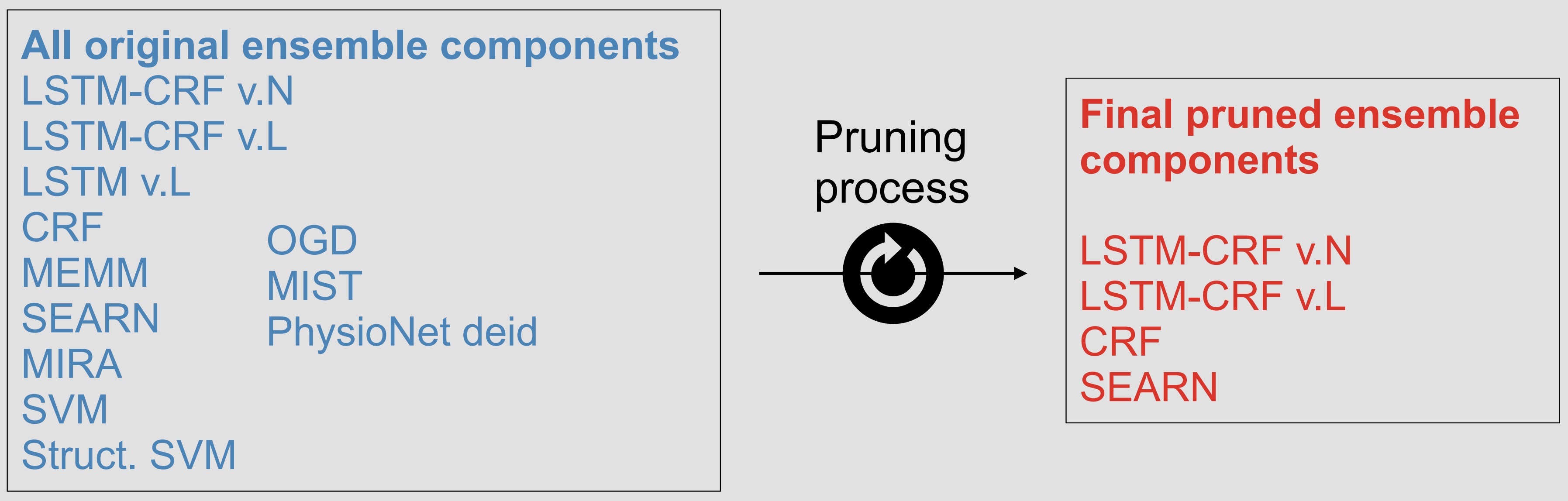
Results

Figure shows the results measured with a pruned voting ensemble, an ensemble with all models, and the best individual de-identification model on the 2014 i2b2[2] test sets.

Note that the y-axis scale in each graph does not start at zero to focus on the value ranges of interest.

The results with voting thresholds ranging from one to six are presented. The precision rates increase as the threshold gets higher and voting pruning (red-colored curve) achieved 97.9% precision.

As shown in the Figure, when the voting threshold was set to two with a pruned ensemble (with 4 members), it achieved the highest F_1 score (95.6%), significantly better than the corresponding ensemble including all members (blue-colored curve) and the best individual model (black-colored line) at the 95% significance level.



All original ensemble components
LSTM-CRF v.N
LSTM-CRF v.L
LSTM v.L
CRF
MEMM
SEARN
MIRA
SVM
Struct. SVM
OGD
MIST
PhysioNet deid

Pruning process

Final pruned ensemble components
LSTM-CRF v.N
LSTM-CRF v.L
CRF
SEARN

References:

1. Kim Y, Heider P, Meystre SM, Ensemble-based Methods to Improve De-identification of Electronic Health Record Narratives. Proceedings of the 2018 AMIA Annual Symposium; 2018.
2. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1. J Biomed Inform. 2015;58:S11–S9.

Contact: kimy@musc.edu

Clinacuity
www.clinacuity.com