Topic Classification of Spoken Inquiries Based on Stacked Generalization

Rafael Torres*, Hiromichi Kawanami*, Tomoko Matsui†, Hiroshi Saruwatari* and Kiyohiro Shikano*
* Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan
E-mail: {rafael-t, kawanami, sawatari, shikano}@is.naist.jp
† Department of Statistical Modeling, The Institute of Statistical Mathematics, Tokyo, Japan
E-mail: tmatsui@ism.ac.jp

***Abstract*—Stacked generalization is a method that allows combining output of multiple classifiers using a second-level classification, minimizing the generalization error of first-level classifiers and achieving greater predictive accuracy. In a previous work, we compared the performance of support vector machine (SVM) with radial basis function (RBF) kernel, prefixspan boosting (pboost) and maximum entropy (ME) in the classification in topics of spoken inquiries in Japanese received by a guidance system. In the present work, we employ a stacked generalization scheme that uses predictions of SVM with RBF kernel, pboost and ME as input for a second-level classification using linear SVM. Experimental results show an improvement in performance from 94.2% to 95.1% in the classification of automatic speech recognition (ASR) 1-best results of adults' inquiries and from 88.3% to 89.2% for children's inquiries, when using stacked generalization in comparison to the individual performance of the first-level classifiers.**

I. Introduction

Stacked generalization, originally proposed by Wolpert in 1992[1], is a method that allows combining output of multiple classifiers using a second-level classification, minimizing the generalization error of first-level classifiers and achieving greater predictive accuracy. Its success arises from its ability to exploit the diversity in the predictions of first-level classifiers.

In a previous work[2], we compared the performance of support vector machine (SVM) with radial basis function (RBF) kernel, prefixspan boosting (pboost) and maximum entropy (ME) in the topic classification of spoken inquiries in Japanese received by a guidance system operating in a real environment. Topic classification in this kind of systems is useful to identify which are users' main information needs and to ease the selection of proper responses to users' inquiries.

The guidance system in mention is the *Takemaru-kun* system[3], which is a real-environment speech-oriented guidance system placed inside the entrance hall of the Ikoma City North Community Center located in the Prefecture of Nara, Japan. The system provides guidance to visitors regarding the center facilities, services, neighboring sightseeing, weather forecast, and news, among other information. The interaction with the system follows a one-question-to-one-response strategy, which fits the purpose of responding simple questions to a large number of users.

In the present work, we employ a stacked generalization scheme that uses predictions of SVM with RBF kernel, pboost and ME as input for a second-level classification using linear SVM. To avoid bias, first-level models are trained using cross-validation; and the predictions that result from these models are used as new data for training a second-level model. As it is shown in experimental results, the proposed stacked generalization scheme can improve the overall predictive accuracy, in comparison to the individual performance of the first-level classifiers.

A. Related Work

In the work of Ting *et al.*[4], the effectiveness of stacked generalization was demonstrated for combining three different learning algorithms: C4.5, Naive Bayes and IB1, using a multi-response linear regression (MLR) algorithm, for the classification of datasets from the UCI repository of machine learning databases. In the work of Sigletos *et al.*[5], stacked generalization was compared against voting, which does not use a second-level classification but takes in consideration the prediction of the majority of the classifiers, concluding that while voting was effective in most of the tested domains, stacked generalization was consistently effective in all the tested domains. In both works it was concluded that using output class probabilities, rather than class predictions, from the first-level classifiers leads to better classification performance.

Sill *et al.*[6] implemented a stacked generalization based technique named feature-weighted linear stacking (FWLS), which was a key component in the solution that awarded them the second place in the Netflix Prize competition carried out in 2009. The objective of the competition was to predict the preferences of customers for various products using the Netflix Prize collaborative filtering dataset. Stacked generalization was also extensively used in the solution of the team that won the first prize, who used a blend of hundreds of different models.

The remainder of the paper is structured as follows: In Section II, the first and second-level classifiers, as well as the proposed stacked generalization scheme are explained. Section III presents the conducted experiments and their results. Finally, Section IV presents the conclusions of the work.

II. Classification with Stacked Generalization

In this section, the first and second-level classifiers as well as the proposed stacked generalization scheme are explained.

A. First-Level Classifiers

In our proposed stacked generalization scheme we use SVM with RBF kernel, pboost and ME as first-level classifiers.

1) Support Vector Machine: Support vector machine (SVM) tries to find optimal hyperplanes in a feature space that maximize the margin of classification of data from two different classes. We used LIBSVM[7] to apply SVM with soft-margin for unbalanced amount of samples, whose primal problem formulation follows the form:

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C_+ \sum_{y_i=1} \xi_i + C_- \sum_{y_i=-1} \xi_i \\ \text{sb.t.} \quad & y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, ..., l \end{aligned} \tag{1}$$

where $\boldsymbol{x}_i \in \mathbb{R}^n, i = 1, ..., l$ indicates a training vector, $y_i \in \{1, -1\}$ a class, and ϕ is the function for mapping the training vectors into feature space. The hyperparameters C_+ and C_- penalize the sum of the slack variable ξ_i for each class, that allows the margin constraints to be slightly violated. By introducing different hyperparameters C_+ and C_-, the unbalanced amount of data problem, in which SVM parameters are not estimated robustly due to unbalanced amount of training vectors for each class, can be dealt with.

We used bag-of-words (BOW) to represent utterances as vectors, and selected a radial basis function (RBF) kernel, which is defined as:

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma||\boldsymbol{x}_i - \boldsymbol{x}_j||^2), \gamma > 0 \tag{2}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represent sample vectors and γ is a hyperparameter of the function.

We used a one-vs-rest approach for multi-class classification, which constructs one binary classifier for each topic. Each classifier is trained with data from a topic, regarded as positive, and the rest of the topics, regarded as negative.

2) PrefixSpan Boosting: Prefixspan boosting (pboost) is a method proposed by Nozowin *et al.*[8]. Pboost implements a generalization of the prefixspan algorithm by Pei *et al.* to find optimal discriminative patterns, and in combination with the linear programming boosting (LPBoost) classifier, it optimizes the classifier and performs feature selection simultaneously.

In pboost, the presence of a single discriminative pattern in a sample, in our case a character sequence that could include gaps, is checked by weak hypotheses, which have the form $h(\boldsymbol{x}; \boldsymbol{s}, \omega)$, where $\boldsymbol{x} \in \{\boldsymbol{x}_i\}$, $\boldsymbol{x}_i \in \mathbb{R}^n, i = 1, ..., l$ is a training vector, $\boldsymbol{s}$ is a character sequence and $\omega \in \Omega, \Omega = \{-1, 1\}$ is a variable that allows the sequence to decide for either class.

The classification function has the form:

$$f(\boldsymbol{x}) = \sum_{(\boldsymbol{s},\omega)\in\bar{\boldsymbol{S}}\times\Omega} \alpha_{\boldsymbol{s},\omega} h(\boldsymbol{x}; \boldsymbol{s}, \omega) \tag{3}$$

where $\alpha_{\boldsymbol{s},\omega}$ is a weight for a character sequence $\boldsymbol{s}$ and parameter ω such that $\sum_{(\boldsymbol{s},\omega)\in\bar{\boldsymbol{S}}\times\Omega} \alpha_{\boldsymbol{s},\omega} = 1$ and $\alpha_{\boldsymbol{s},\omega} \geq 0$, which indicates the discriminative importance of a character sequence.

To deal with the unbalance between positive and negative samples, we used an extended version of the method that allows to implement soft-margin for unbalanced amount of samples. The pboost primal problem then takes this form:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\xi,\rho} \quad & -\rho + D_+ \sum_{y_i=1} \xi_i + D_- \sum_{y_i=-1} \xi_i \\ \text{sb.t.} \quad & \sum_{(\boldsymbol{s},\omega)\in\bar{\boldsymbol{S}}\times\Omega} y_i \alpha_{\boldsymbol{s},\omega} h(\boldsymbol{x}_i; \boldsymbol{s}, \omega) + \xi_i \geq \rho, i = 1, ..., l \\ & \sum_{(\boldsymbol{s},\omega)\in\bar{\boldsymbol{S}}\times\Omega} \alpha_{\boldsymbol{s},\omega} = 1, \boldsymbol{\alpha} \geq 0, \boldsymbol{\xi} \geq 0 \end{aligned} \tag{4}$$

where $\boldsymbol{x}_i \in \mathbb{R}^n, i = 1, ..., l$ indicates a training vector, $y_i \in \{1, -1\}$ a class, ρ is the soft margin separating negative from positive samples, and $D = \frac{1}{\nu\ell}$ and $\nu \in (0, 1)$ is a hyperparameter controlling the cost of misclassification, which in this case is separated into D_+ and D_-, penalizing the sum of the slack variable ξ_i for each class, that allows the margin constraints to be slightly violated. As in SVM, by introducing different hyperparameters ν_+ and ν_-, we can deal with the unbalanced amount of data problem.

Here, we also used a one-vs-rest approach for multi-class classification.

3) Maximum Entropy: Maximum entropy (ME) is a technique for estimating probability distributions from data, which has been widely used in natural language tasks, including speech classification, where it has shown to outperform other conventional statistical classifiers[9].

As expressed in [9], given an utterance consisting of the character sequence c_1^N, the objective of the classifier is to provide the most likely class label $\hat{k}$ from a set of labels K:

$$\hat{k} = \underset{k\in K}{\operatorname{argmax}}\ p(k|c_1^N) \tag{5}$$

where the ME paradigm expresses the probability $p(k|c_1^N)$ as:

$$p(k|c_1^N) = \frac{\exp\left[\sum_c N(c)\log\alpha(k|c)\right]}{\sum_{k'}\exp\left[\sum_c N(c)\log\alpha(k|c)\right]}. \tag{6}$$

Ignoring the terms that are constant with respect to k yields:

$$\hat{k} = \underset{k\in K}{\operatorname{argmax}} \sum_c N(c)\log\alpha(k|c) \tag{7}$$

where $N(c)$ is the frequency of a character or character sequence in an utterance, and $\alpha(k|c)$ with $\alpha(k|c) \geq 0$ and $\sum_k \alpha(k|c) = 1$ are parameters that depend on a class k and a character or character sequence c.

We applied ME with the package maxent Ver.2.11[10] using the ME model with inequality constraints. The parameters are estimated using the L-BFGS-B algorithm, which is a limited-memory algorithm for solving large nonlinear optimization problems subject to simple bounds on the variables.

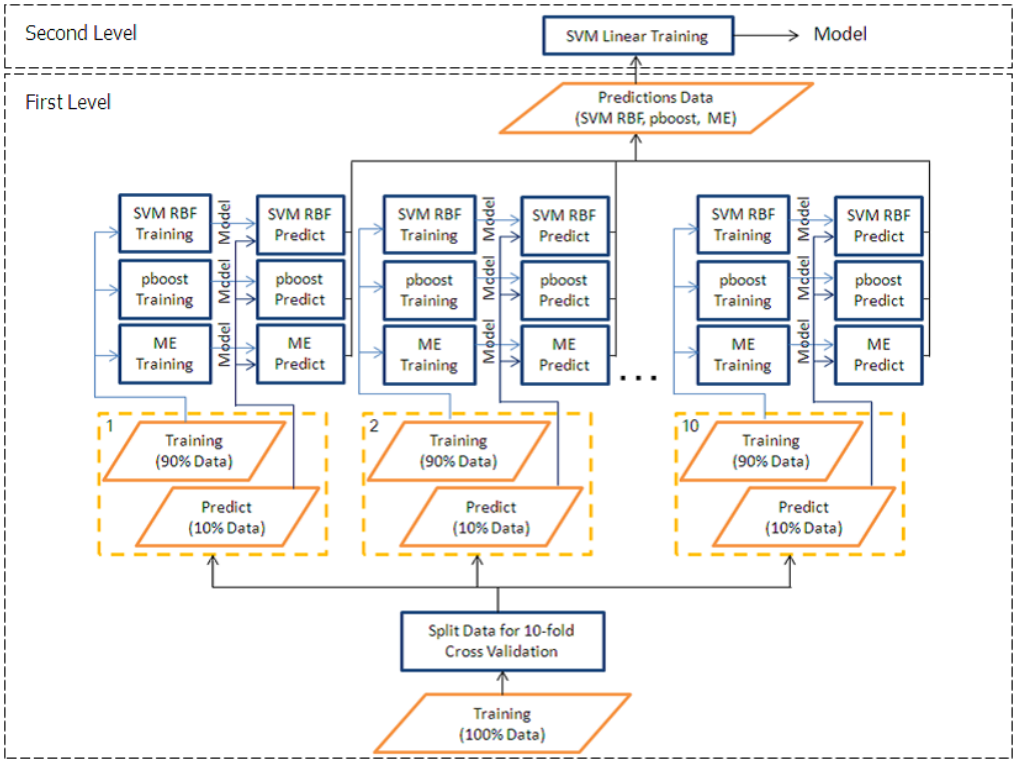

Fig. 1. Stacked generalization scheme

B. Second-Level Classifier

We used linear SVM as second-level classifier, as described in Section II-A1 but differing in the kernel function:

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j \tag{8}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represent sample vectors.

C. Stacked Generalization Scheme

Stacked generalization combines output of multiple classifiers using a second-level classification. The first step in the stacked generalization training is to collect the predictions of each first-level classifier to create a new set of data, containing the true classification and the predictions of each one of the classifiers for each one of the samples of the original dataset.

To avoid bias, the first-level models should be trained excluding the samples we want to predict, which can be achieved by using cross-validation. In our implementation we used 10-fold cross-validation. The second step is to use the predictions of the first-level classifiers as new data for training a second-level model.

The scheme of the stacked generalization training that we implemented is presented in Figure 1. The original training set is split in ten parts to implement 10-fold cross validation, and each topic is represented in the same proportion on each one of the resultant sets. Then, we train the first-level models using these sets, and obtain the predictions of SVM RBF, pboost and ME for each one of the samples. Finally, we use these predictions to train the second-level model using SVM linear. Each one of the sample feature vectors used to train the second-level model contains the real topic of the sample and the predictions for each topic. SVM RBF and pboost give 1 or 0 depending if the sample is positive in a certain topic, and ME gives a probability for each topic.

III. Experiments

We compared the performance of the stacked generalization scheme described above with the individual performance of SVM RBF, pboost and ME, in the classification in topics of ASR results of utterances in Japanese received by the speech-oriented guidance system *Takemaru-kun*. We used character unigrams, bigrams and trigrams as features for the training of first-level models, as it was shown to improve classification accuracy in comparison to words[2]. Optimal hyper-parameter values for SVM and pboost were obtained experimentally using a grid search strategy, and were set a posteriori. The experiments and obtained results are detailed below.

A. Characteristics of the Datasets

The data used in the experiments were valid utterances from adults and children, collected by *Takemaru-kun* from Nov. 2002 to Oct. 2004. Julius Ver.3.5.3 was used as ASR engine. Acoustic models (AMs) and language models (LMs) were separately prepared for adults and children. The AMs were trained using the Japanese Newspaper Article Sentences (JNAS) database, adapting them with the samples collected by the system. The LMs were constructed using the transcriptions of the samples. Samples corresponding to the month of Aug.

2003 were used for the test sets and were not included in the training sets. For these experiments we selected the 15 topics with most training samples, e.g. greeting-start, info-facility, info-weather and others. We conducted experiments with transcriptions and ASR 1-best results. Table I shows the word recognition accuracy of the ASR engine for the datasets, as well as the amount of samples and the sizes of the vocabularies, which were composed by character unigrams, bigrams and trigrams.

B. Experiment Results

To test the second-level models, we obtained predictions of the test set samples by classifying them using first-level models trained with the entire training datasets. Those predictions were used as test data for the second-level model. The classification performance of the second-level model is the classification performance of the stacked generalization.

The classification performance of the methods was evaluated using the F-measure, which was calculated individually for each topic and it was averaged by frequency of samples in the topics. Figures 2 and 3 present the performance of SVM RBF, pboost and ME individually, and combined using the stacked generalization scheme, for adults and children.

In both cases, the classification performance obtained by using stacked generalization is comparatively higher than the individual performance of the methods. When classifying ASR 1-best results, for adults' data SVM RBF and pboost individually presented better performance than ME, with 94.2%, and the stacked generalization yielded to a performance of 95.1%. For children's data ME individually performed better than SVM RBF and pboost, with 88.3%, and the stacked generalization yielded to 89.2%.

Additional experiments combining several SVM RBF classifiers trained with different hyperparameters, with pboost and ME using stacked generalization, did not improve classification performance. Experiments excluding pboost yielded to a decrease in the classification performance of stacked generalization, in spite of its lower classification accuracy for children's data.

IV. Conclusions

This work described a stacked generalization scheme to combine predictions of SVM, pboost and ME, using a second-level classification with linear SVM. The resultant performance improvement with the stacked generalization scheme is relatively small, which is reasonable as individual classifiers presented similar classification errors; however, experimental results show that it can improve the overall predictive accuracy, in comparison to the individual performance of the first-level classifiers.

TABLE I
ASR word recognition accuracy, sample amount and vocabulary size per dataset

	Adult		Child	
	Train	Test	Train	Test
ASR acc. (%)	85.66	85.10	66.81	67.18
Sample amount	14431	792	43494	3738
Vocab. size (transcriptions)	12481		32108	
Vocab. size (ASR 1-best)	13287		37248	

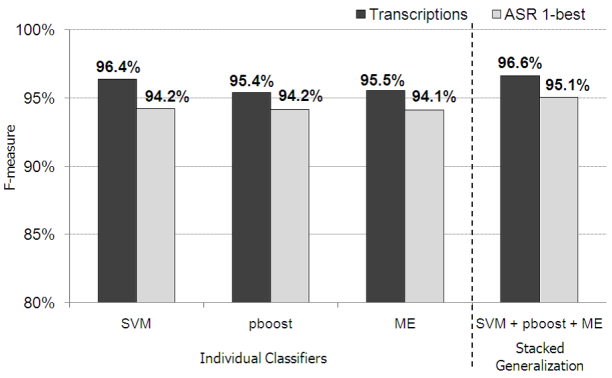

Fig. 2. F-measure per method for adults' inquiries

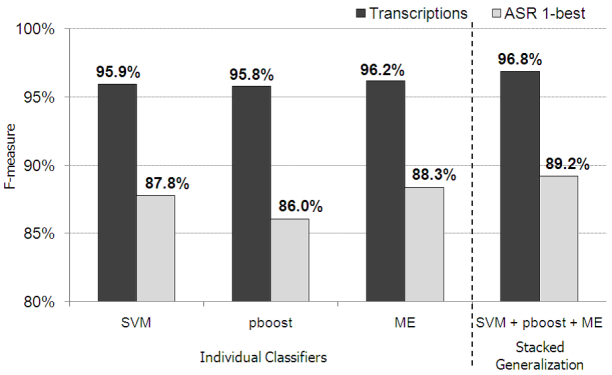

Fig. 3. F-measure per method for children's inquiries

References

[1] D.H. Wolpert, "Stacked Generalization," *Neural Networks*, Vol.5(2), pp.241-260, 1992.

[2] R. Torres, S. Takeuchi, H. Kawanami, T. Matsui, H. Saruwatari, K. Shikano, "Comparison of Methods for Topic Classification in a Speech-Oriented Guidance System," *In Proc. of Interspeech 2010*, pp.1261-1264, 2010.

[3] R. Nisimura, A. Lee, H. Saruwatari, K. Shikano, "Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability," *In Proc. of ICASSP 2004*, Vol.1, pp.433-436, 2004.

[4] K.M. Ting, I.H. Witten, "Issues in Stacked Generalization," *Journal of Artificial Intelligence Research*, Vol.10, pp.271-289, 1999.

[5] G. Sigletos, G. Paliouras, C.D. Spyropoulos, M. Hatzopoulos, "Combining Information Extraction Systems Using Voting and Stacked Generalization," *Journal of Machine Learning Research*, Vol.6, pp.1751-1782, 2005.

[6] J. Sill, G. Takacs, L. Mackey, D. Lin, "Feature-Weighted Linear Stacking," *CoRR*, arXiv:0911.0460, 2009.

[7] C. Chang, C. Lin, "LIBSVM: a Library for Support Vector Machines," 2001. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm

[8] S. Nowozin, G. Bakir, K. Tsuda, "Discriminative Subsequence Mining for Action Classification," *In Proc. of ICCV 2007*, Software available at http://www.kyb.mpg.de/bs/people/nowozin/pboost

[9] K. Evanini, D. Suendermann, R. Pieraccini, "Call Classification for Automated Troubleshooting on Large Corpora," *In Proc. of ASRU 2007*, pp.207-212, 2007.

[10] Maximum Entropy Modeling Package. http://mastarpj.nict.go.jp/ mutiyama/software.html