A Simple Ensemble Technique

Team: ISMLL
Members: Krisztian Buza and Lars Schmidt-Thieme
buza@ismll.de, schmidt-thieme@ismll.de
Information Systems and Machine Learning Lab
University of Hildesheim, Germany, European Union

November 24, 2009

Abstract

In this paper we present our solution for the *AusDM Analytic Challenge 2009*. We applied a simple approach based on a smart variable selection technique. The basic idea is looking at the pairs of models and searching for those pairs where the errors of the two models compensate each other.

As the final results of the challenge are unknown at the time of writing this paper, we can only report preliminary results on the small data set: according to this, despite its simplicity, our technique is less than 0.1% worse than the currently best method[1].

1 Introduction

The topic of the *AusDM Analytic Challenge 2009* was ensembling: "Ensembling, Blending, Committee of Experts are various terms used for the process of improving predictive accuracy by combining models built with different algorithms, or the same algorithm but with different parameter settings."[2] This technique is frequently used to improve predictive models, see e.g. [4, 3, 2]. On the one hand some fundamental reasons are known, why ensembles work better than single models (these reasons are described for example in [1]), on the other hand, how ensembles are "actually achieved in practice maybe somewhat arbitrary. One of the drawbacks in researching the problem is that you first have to generate a lot of models before you can even start. There have been numerous predictive modelling competitions that could potentially provide good data sets for such research - many models built by many experts using many techniques. The Netflix Prize is one such competition that has been on going for nearly 3 years."

[1] 17th November 2009, 11:40 GTM

[2] All the cited text in the Introduction is from http://www.tiberius.biz/ausdm09/

In the Netflix challenge the task was to predict how users rate movies on a 1 to 5 integer scale (5=best, 1=worst). Participants of the challenge delivered real numbers as predictions and the RMSE (Root Mean Squared Error) between the predictions and the actual ratings was used as evaluation metric.

The Netflix challenge "recently finished, and the eventual winners were an amalgamation of several teams that somehow combined their individual model predictions. Over 1,000 sets of predictions have been provided by the two leading teams (who actually ended up with the same score), The Ensemble and BellKor's Pragmatic Chaos."

In the AusDM Analytic Challenge 2009 the task was to build an ensemble over these provided predicitions. There were two subtasks "one to develop a method to predict a continuous value" (RMSE-task) "and the other to predict a binary value" (AUC-task). We only participated in the RMSE-subtask.

For the participants of the challenge 6 datasets were provided: for both tasks there were small, medium and large datasets. During the challenge feedback was given on the small dataset, but the final results were determined based on the performance on the medium and large datasets (for which no feedback was provided during the challenge).

2 A Simple Ensemble Technique

Ensembles work better than a single predictive model in those situations when different predictive models have different error characteristics and their errors can compensate each other. We build our simple ensembling technique based on this observation. In fact, we applied a variant of stacking with linear regression as meta learner.

For simplicity we will describe our technique in context of linear regression (as meta learner) and RMSE as evaluation score, but the same technique work with arbitrary classification or regression models (like SVMs, Bayesian Networks, Decision Tree, etc.) as meta learner and various evaluation scores (like accuracy, AUC, etc.), thus our technique is quite general in spite of the current description which is very specific. The only assumption we make, is that *each predictive model* assisting in the ensemble *delivers a prediction for the target.*

While learning, we use a technique similar to 10 fold crossvalidation: first we devide the train data into 10 splits, which are numbered 0, 1, 2..., 9. In the 1st fold the splits 0, 1, 2, 3 and 4 serve as *basic train set* and the rest as *internal evaluation set.* In general, in the kth fold, the *basic training set* contains the splits k mod 10, $(k + 1)$ mod 10, $(k + 2)$ mod 10, $(k + 3)$ mod 10 and $(k + 4)$ mod 10, and the *internal evaluation set* contains the rest.

What we describe from now on, is done for each fold. Our ensembling technique looks at the pairs of models and searches for those pairs where the errors of the two models compensate each other. For the simplicity of the description, we introduce the *model-pair graph.* The *model-pair graph* is a weighted, undirected, graph. Each vertex v of this graph corresponds to one of the models that assist in the ensemble. All the vertexes are connected, i.e. the graph is a complete graph. Each edge $\{v_i, v_j\}$ has a weight reflecting how "good" is the *combination* of the models v_i and v_j, that is how well they compensate each other's errors. The weight of $\{v_i, v_j\}$ is determined on the *basic train set*: we regard the average of the outputs of the models v_i and v_j as prediction for the target variable, and we calculate the RMSE between this average and the actual value of the target. This RMSE score will be the weight of the edge $\{v_i, v_j\}$.

We process the edges in order of their scores, beginning with the best edge. (As in case of RMSE the smaller values indicate the better predictions, we process the edges in *ascending* order with respect to their weights.) Let M denote a set of models, initially M is the empty set. Let s denote the estimated quality of our ensemble based on the models in M. Initially s is the worst possible quality score, i.e. $s = +\infty$ (positive infinity), as in case of RMSE scores. Let $\epsilon = 0.25$.

For each edge $\{v_i, v_j\}$ the followings have to be done:

1. If both $v_i \in M$ and $v_j \in M$, then proceed for the next edge,
2. else
   (a) $M' = M \cup \{v_i, v_j\}$.
   (b) Train a multivariate *linear regression* over the outputs of the models contained in M' using the *basic train set*, and evaluate it on the *internal evaluation set.* Let s' denote the evaluation score. If s' is better than s at least by ϵ (i.e. if $(s' + \epsilon) < s$ for RMSE), than $M \leftarrow M'$ and $s \leftarrow s'$.
   (c) Proceed for the next edge.

This way for each fold we select a set of models M, or being more exact: we get M_0, M_1,..., M_9 for the folds 0,1,...,9. Let N denote the set of such models that are contained at least $n = 4$ times among the selected models, i.e. N is a set of such models that are contained in at least $n = 4$ sets among the sets M_0,M_1,...,M_9.

Finally we train a *linear regression* over the output of the models in N on the whole training set and apply this for the unlabeled data. We built our implementation on the linear regression of WEKA[3] software package [5].

The hyperparameters (ϵ and n) are to be learned on a hold-out subset of the train data that is disjoint both from the basic train set and from the internal evaluation set.

3 Preliminary Evaluation and Outlook

As the challenge is not yet finished at the time of writing this paper, we only report preliminary results that were achived on the small data set, where feedback was given to the participants. On the score set of the small data set, the currently[4] best method (Optibrebs) achieved RMSE of 877.907. Our technique

[3] http://www.cs.waikato.ac.nz/~ml/index.html

[4] 17th November 2009, 11:40 GTM

Table 1: Experiments

Method	Performance (RMSE)
Optibrebs (currently best)	877.907
Simple Ensemble Technique (Our method)	878.612
Average Top 10 Experts (Baseline)	884.359
Best Expert (Baseline)	888.32
Average All Experts (Baseline)	892.77
Worst Expert (Baseline)	994.118

had RMSE of 878.612, which is less than 0.1 % worse, than the RMSE of the best model, whereas the best baseline method (average top 10 experts) achived RMSE of 884.359. The performances of our technique and the baseline methods and the currently best method are summarized in Table 1.

The simple technique presented in this paper can be improved in several directions, for example: (i) one can consider to check not only pairs, but also triples of variables, (ii) new edge weighting strategies can be introduced, (iii) the technique can simply be applied for other evaluation measures than RMSE. As future work, we would also like to explore how general this simple ensembling technique can be applied, as a first step on can measure the performance on other data sets.

Acknowledgements. This work was co-funded by the EC FP7 project MyMedia under the grant agreement no. 215006. For inquiries please contact info@mymediaproject.org.

References

[1] T. G. Dietterich (2000): *Ensemble Methods in Machine Learning*, MCS 2000, LNCS 1857, pp. 1-15., Springer

[2] Y. Peng (2006): *A novel ensemble machine learning for robust microarray data classification*, Computers in Biology and Medicine, Volume 36, Issue 6, Pages 553-573

[3] C. Preisach, L. Schmidt-Thieme (2008): *Ensembles of Relational Classifiers* Knowledge and Information Systems Journal 14(3)

[4] A. C. Tan, D. Gilbert (2003): *Ensemble machine learning on gene expression data for cancer classification*, New Zealand Bioinformatics Conference

[5] Ian H. Witten, Eibe Frank (2005): *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)* Morgan Kaufmann, ISBN 0-12-088407-0