Ensemble Methods for Personality Recognition

Ben Verhoeven and **Walter Daelemans** and **Tom De Smedt**
CLiPS, University of Antwerp
Prinsstraat 13 (L), 2000 Antwerp, Belgium

Abstract

An important bottleneck in the development of accurate and robust personality recognition systems based on supervised machine learning, is the limited availability of training data, and the high cost involved in collecting it. In this paper, we report on a proof of concept of using ensemble learning as a way to alleviate the data acquisition problem. The approach allows the use of information from datasets from different genres, personality classification systems and even different languages in the construction of a classifier, thereby improving its performance. In the exploratory research described here, we indeed observe the expected positive effects.

Introduction

In personality recognition, the goal is to assign a personality profile to the author of a text. Possible applications of this task include social network analysis and user modeling in conversational systems. Currently, this task is most effectively handled using supervised machine learning methods. See for example (Mairesse et al. 2007; Luyckx and Daelemans 2008; Noecker, Ryan, and Juola 2013) and references therein. Training texts are collected and linked to personality profiles of the authors, resulting from personality tests taken by the authors or from judgements of experts. Collecting these data is a costly process, resulting in relatively little available training data. In addition, available data is distributed over different genres (essays, stream of consciousness text, social network text), using different personality typing systems (MBTI (Briggs Myers and Myers 1980) or Big Five (Goldberg 1990)), and in different languages.

Ensemble methods (Seni and Elder 2010) are an approach in Machine Learning that has been successful in improving the accuracy of systems by combining the predictions of different component classifiers. Apart from approaches based on creating different classifiers from different subsets of the data or different subsets of the features (bagging, random forests) or methods based on combining weak learners (boosting), it is also possible to create meta-learners (sometimes called stacked learning) that learn to make predictions on the basis of input features with as additional input the outputs of several component classifiers on that input. The key advantage of this approach is that the component classifiers need not even use the same class inventory as that of the task, or may be trained on data from different genres. For example, in (Zavrel and Daelemans 2000), it is shown that a wide variety of existing resources can be integrated in a meta-learner for learning a tagger with a new part-of-speech tagset, and (Van Halteren, Daelemans, and Zavrel 2001) shows significant improvements in part-of-speech tagging accuracy using this approach. In this paper, we will apply this relatively rarely used variant of ensemble methods to the personality recognition task.

Our meta-learning approach to personality recognition was done in the context of the data provided in the shared task of the workshop on computational personality recognition[1]. Using the Facebook dataset as target recognition task, we show that classifiers trained on the different personality traits of the essay data (a different genre from the Facebook data) improve performance in an ensemble learning set-up. In this case, the essay data are in the form of a stream of consciousness.

In the following sections, we describe our approach, the data used, and our modeling decisions. We show our results, and we conclude by analyzing the strengths and shortcomings of the approach and possible extensions.

Data and Approach

In this exploratory research, we worked with a 60%-40% split of the Facebook data provided, which we consider here the target task. These data include anonymized authors, status updates in text, gold standard labels (both classes and scores) and a number of social network measures. Anonymization was done by replacing proper names of persons not belonging to a 'famous names' list by a *PROPNAME*. The Big Five labels are used as class symbols: EXT (extraversion), NEU (neuroticity), AGR (agreeableness), CON (conscientiousness), and OPN (openness).

More extensive cross-validation experiments (requiring an embedded cross-validation loop within the first level cross-validation loop, as two levels of classifiers are used) are forthcoming. The current paper is intended as a proof of concept.

[1] http://mypersonality.org/wiki

In this paper we don't focus on the optimization of document representations (which features to use) nor on optimization of algorithm parameters. We are not after optimal performance, but after insight into whether the proposed information combination approach works or not. We use a standard SVM (SMO as implemented in the WEKA[2] tool, (Hall et al. 2009)) with default parameter settings for training of all classifiers, and the 2000 most frequent character trigrams in the training data as document representation in all experiments.

We did five meta-learning experiments with the Facebook data, one for each personality trait. In each of these experiments, the ensemble (meta) learner did a ten-fold cross-validation experiment on the held-out test data, with as training input data the Facebook document vectors (2000 most frequent character trigrams), and as additional input the outputs of the following ten component classifiers.

- Facebook: The output of five classifiers based on a hold-out experiment on the Facebook training and test data (one for each personality trait).
- Essays: The output of five classifiers (one for each personality trait) trained on the complete essay data. This data is completely independent from the Facebook data, so all data can be used.

The meta-learner therefore gets as input 2000 trigram features and 10 predicted classes by component classifiers, and is trained and tested on the Facebook test data in a ten-fold cross-validation experiment. We will go into more detail for each of the steps in the methodology before turning to the results. Figure 1 provides an overview of the general set-up.

The original Facebook dataset (almost 10,000 instances) was simplified by concatenating posts with the same personality type (all five traits identical) into larger instances of 20 posts (or less when fewer posts were left over). We first randomized the data to avoid posts from the same author being collected in the same bins. These larger instances contain better estimations of trigram distributions. Trigrams were computed at post level (i.e. no trigrams cross post boundaries within a bin). Our resulting Facebook data then consists of 509 instances. The Facebook classifiers were trained on 60% of the Facebook data, using the 2000 most frequent trigrams as document representation, and with default SMO parameter settings.

For the Essays classifier we kept to the instances as collected by Pennebaker and used in (Mairesse et al. 2007), in the format provided for the shared task. We trained five classifiers, one for each trait, using the 2000 most frequent trigrams as document representation, on the complete essays dataset, with default SMO parameter settings.

Tables 1 and 2 show the results of the component classifiers (this time using ten-fold cross-validation) on their own data. This gives an idea about the accuracy of these classifiers, trained and evaluated within the same genre. As mentioned before, there has been no effort to optimize document representation and algorithm parameters. In an ensemble method, each of the component classifiers should be above chance, and optimization on the 'native' data may lead to overfitting and lower results when used as part of an ensemble. The baseline in these results is the weighted random baseline (WRB, sum of the square of the probabilities of the classes in the training data). For the essay data, all trait classifiers are above baseline except AGR, for the Facebook data all are above baseline. In general (with the features and machine learning algorithm chosen), the Facebook data seems to be easier than the essay data.

[2]http://www.cs.waikato.ac.nz/ml/weka/

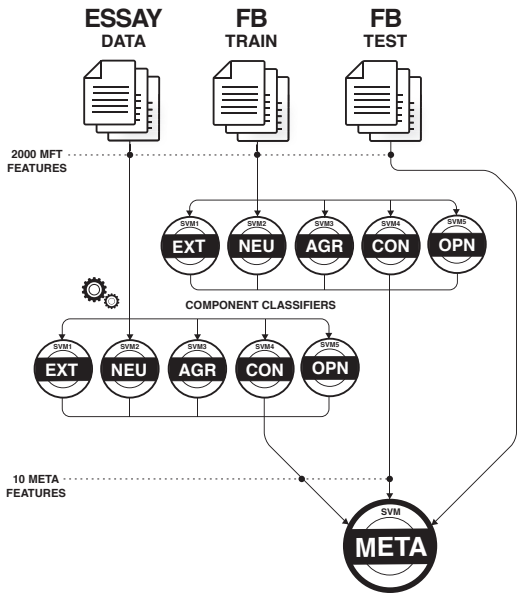

Figure 1: Ensemble set-up (mft is most frequent character trigrams).

Trait	Precision	Recall	F-Score	WRB
EXT	0.53	0.53	0.53	0.50
NEU	0.54	0.54	0.54	0.50
AGR	0.50	0.50	0.50	0.50
CON	0.54	0.54	0.54	0.50
OPN	0.56	0.56	0.56	0.50

Table 1: Results of the component classifiers on essay data using ten-fold cross-validation. Scores are given for accuracy, precision, recall, and F-score, WRB is the Weighted Random Baseline.

Results

Since we have shown that the intended component classifiers for our ensemble score above baseline (apart from one) , we proceed with our meta-learning approach. Table 3 shows the results of the ensemble experiments. These results should

Trait	Precision	Recall	F-Score	WRB
EXT	0.66	0.67	0.66	0.51
NEU	0.61	0.62	0.62	0.53
AGR	0.62	0.62	0.62	0.50
CON	0.75	0.74	0.74	0.50
OPN	0.74	0.75	0.74	0.59

Table 2: Results of the component classifiers on Facebook data using ten-fold cross-validation. Scores are given for accuracy, precision, recall, and F-score, WRB is the Weighted Random Baseline.

be compared to those in Table 4 that are based on a ten-fold cross-validation experiment on the Facebook test data without the information from other component classifiers. For each personality trait, the ensemble improves upon the 'normal' system (except AGR which stays the same), and in two cases (OPN and CON) in a statistically significant way ($p < 0.05$). Statistical significance of differences was computed using an approximate randomization approach; a non-parametric test suitable for F-scores (Noreen 1989)[3]. These results show that the approach works, and that an ensemble approach is a possible way to integrate information from a dataset representing one genre into a model trained for another genre.

Trait	Precision	Recall	F-Score
EXT	0.79	0.79	0.79
NEU	0.71	0.72	0.70
AGR	0.67	0.68	0.67
CON	0.72	0.72	**0.72**
OPN	0.87	0.87	**0.86**

Table 3: Results of the ensemble classifier on the Facebook test partition. Ten-fold cross-validation on test split with 2000 most frequent trigrams and 10 component classifier outputs. Scores are given for precision, recall, and F-score. Statistically significant systems with normal approach are in bold.

Trait	Precision	Recall	F-Score
EXT	0.76	0.77	0.76
NEU	0.67	0.68	0.66
AGR	0.67	0.68	0.67
CON	0.66	0.66	0.66
OPN	0.82	0.83	0.82

Table 4: Results of the single classifier on the Facebook test partition. Ten-fold cross-validation on test split with 2000 most frequent trigrams only. Scores are given for precision, recall, and F-score.

[3]A script implementing the approach, art.py, can be found at http://www.clips.ua.ac.be/scripts/art

An obvious alternative approach to using the essay data together with the Facebook training data would be to simply combine the two datasets and test on the Facebook test data. These results can be found in Table 5. As can be seen there, the results of the ensemble are better, and more interestingly, performance even degrades compared to the results in Table 4.

Trait	Precision	Recall	F-Score
EXT	0.70	0.70	0.70
NEU	0.63	0.64	0.63
AGR	0.62	0.62	0.62
CON	0.58	0.58	0.58
OPN	0.75	0.72	0.73

Table 5: Results of the single classifier on the Facebook test partition. Trained on essays data plus Facebook train partition. Scores are given for precision, recall, and F-score.

Discussion and Conclusion

We consider the results on the Facebook train-test split as a proof of concept that an ensemble method of the meta-learning type can improve the performance of a classifier by using data from a different genre better than the combination of the training data from the different genres. This approach can in principle be extended with other component classifiers from other genres, with other class systems, or even from other languages (through machine translation). The meta-learning approach gets as information both the original input feature vectors and the outputs of the component classifiers. This makes it possible for the ensemble to learn to 'trust' outputs of specific classifiers more than that of others depending on the type of inputs.

However, Table 6 shows that at least for these data, the meta-learner doesn't make use of this capability. Actually, without the character n-gram features as input (so only with the ten component classifier outputs), the meta-learner produces even better results. This is possibly due to the low informative value (given our features) of the essay data, and the relative importance of the Facebook data (including the predictions for other traits than the one being evaluated).

Trait	Precision	Recall	F-Score
EXT	0.74	0.74	0.74
NEU	0.68	0.69	0.68
AGR	0.70	0.70	0.70
CON	0.74	0.74	0.74
OPN	0.85	0.85	0.85

Table 6: Results of the ensemble classifier on the Facebook train-test partition without ngrams as input. Scores are given for precision, recall, and F-score.

One other negative result needs reporting on. We also applied the approach the other way round, taking the essay data personality recognition as the central task, and using output

of classifiers trained on the Facebook data as part of the out-of-genre ensemble. Although some small increases in accuracy could be found, none were statistically significant, and performance even deteriorated for other traits. Further analysis and experimentation is needed to find out more about the reasons for this. Again, the essay data turns out to be a very hard dataset compared to the Facebook data, at least with the features and learning method chosen in this paper.

In conclusion, we have shown that an ensemble method based on meta-learning allows the productive use of out-of-genre data in a more effective way than simpler approaches. This approach has large potential because in principle any relevant data (from other domains or registers, from other languages, and using other class systems) can be used to achieve better performance. However, in this paper we have only been able to show the proof of concept, and our positive results are counterbalanced by negative results. In future research we will investigate these issues further in a complete cross-validation set-up and using additional information sources and error analysis to provide deeper insight into the conditions under which this approach is useful.

Acknowledgements

This research was funded by the Flemish government agency for Innovation by Science and Technology (IWT) through SBO (Strategic Basic Research) project AMiCA.

References

Briggs Myers, I., and Myers, P. B. 1980. *Gifts differing: Understanding personality type*. Mountain View, CA: Davies-Black Publishing.

Goldberg, L. R. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology* 59(6):1216–1229.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter* 11(1):10–18.

Luyckx, K., and Daelemans, W. 2008. Personae: a corpus for author and personality prediction from text. In *Proceedings of the 6th International Conference on Language Resources and Evaluation.* Marrakech, Morocco: European Language Resources Association.

Mairesse, F.; Walker, M. A.; Mehl, M. R.; and Moore, R. K. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research* 30(1):457–500.

Noecker, J.; Ryan, M.; and Juola, P. 2013. Psychological profiling through textual analysis. *Literary and Linguistic Computing*.

Noreen, E. W. 1989. *Computer-Intensive Methods for Testing Hypotheses : An Introduction*. Wiley-Interscience.

Seni, G., and Elder, J. F. 2010. Ensemble methods in data mining: improving accuracy through combining predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery* 2(1):1–126.

Van Halteren, H.; Daelemans, W.; and Zavrel, J. 2001. Improving accuracy in word class tagging through the combination of machine learning systems. *Computational linguistics* 27(2):199–229.

Zavrel, J., and Daelemans, W. 2000. Bootstrapping a tagged corpus through combination of existing heterogeneous taggers. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, 17–20. Athens, Greece: European Language Resources Association.