Meta^2-Features: Providing Meta-Learners More Information

Matthias Reif, Faisal Shafait, and Andreas Dengel

German Research Center for Artificial Intelligence,
Trippstadter Str. 122, 67663 Kaiserslautern, Germany
{matthias.reif,faisal.shafait,andreas.dengel}@dfki.de

Abstract. Meta-features are used to describe properties and characteristics of datasets and construct the feature space for meta-learning. Many of the different meta-features are defined for single variables and, therefore, are computed per feature of the dataset. Since datasets contain different numbers of features but meta-learning requires feature vectors of the same size, such measures are typically simply averaged over all columns.

In this paper, we present an approach of preserving more information of such meta-features while producing a feature vector with a fixed size. An additional level of features are extracted from the meta-features.

1 Introduction

Meta-features are a well known concept in the meta-learning domain. They are measures calculated on a dataset in order to describe its properties and characteristics. Meta-features construct the feature space in which each dataset is represented as a point. Multiple datasets as points within this feature space are used as training data for meta-learning: Knowledge about these datasets (e.g. the best performing classifier) is used to infer knowledge about a new dataset, e.g. predicting the best performing classifier. Statistical pattern recognition methods are applied to create a model that is able to make the desired prediction by applying the model on the meta-features of a new dataset.

Using meta-features, various meta-learning tasks have been developed. The most prominent meta-learning problem is model or algorithm selection, that has been addressed by applying classification [1], regression [11], and ranking [4], but also parameter optimization can be tackled by meta-learning [10].

Commonly used types of meta-features are statistical and information-theoretic measures [7, 5, 6, 12]. Two statistical meta-features that are often used are the skewness and the kurtosis. The entropy and the joint-entropy are two simple examples of information-theoretic meta-features. Other types of meta-features are landmarking [9, 2] and model-based features [3, 8].

An issue of many statistical and information-theoretic meta-features is that they are defined on single features of the dataset. Computing such measures for all features leads to a different number of values for datasets with different

S. Wölfl (Ed.): Poster and Demo Track of the 35th German Conference on
Artificial Intelligence (KI-2012), pp. 74-77, 2012.

numbers of features. Additionally, the meta-feature vectors for datasets with the same number of features are not useful because the order of the features have an influence on the meta-features but obviously not on the characteristics of the dataset. If the meta-features are calculated per feature and additionally per class [7, 12], this issue is further strengthened. Therefore, such meta-features are typically averaged [7, 5, 12]. This leads to a meta-feature vector with the same size and semantics for differently sized datasets, but also to an high loss of information.

Spiliopoulou et al. [13] proposed to use the minimum, maximum, and standard deviation of the number of examples per class, the number of distinct values of the attributes, and the number of missing values of the attributes in addition to the average value. Using the minimum, maximum, and the standard deviation in addition increases the amount of information about the dataset. In our paper, we go one step further and propose to use meta-features of meta-features in order to keep as much information as possible. This can be seen as an generalization of the meta-features used by Spiliopoulou et al.

2 Approach

The proposed approach is divided into two steps. First, the per-feature meta-features are calculated for each feature. They are collected and construct an intermediate dataset where each column is a meta-feature (e.g. skewness) and each line is a feature of the original dataset. The value of a cell is the meta-feature value of the original feature. While the number of features of this intermediate dataset is the number of meta-features used and, therefore, the same for each original dataset, the number of instances is differently.

In the next step, meta-features of this intermediate dataset are calculated. This might be a subset of the meta-features of the previous step. For example, the entropy of the kurtosis values of the features might be computed. This step leads to a single vector with the same length also for original datasets with different number of features.

The two steps of the presented approach are illustrated in Figure 1: The meta-features skewness, kurtosis, entropy, and mutual information are calculated for each of the three features of the original dataset. These values construct the intermediate dataset with four columns and three rows. Afterwards, the minimum, maximum, mean, standard deviation, skewness, kurtosis, and entropy are calculated for each of the previous meta-features. This leads to $4 \times 7 = 28$ meta^2-features. Since the mean is also calculated, the set of meta^2-features also contains the traditional meta-features and the measures of Spiliopoulou et al. [13]. Of course, other meta-features such as landmarking can be added to the vector as well.

Figure 2 shows the distribution of eight features for three artificial datasets as an illustrating example. All three datasets have a similar mean skewness of about 1.37. However, meta^2-features are able to describe the difference of the datasets:

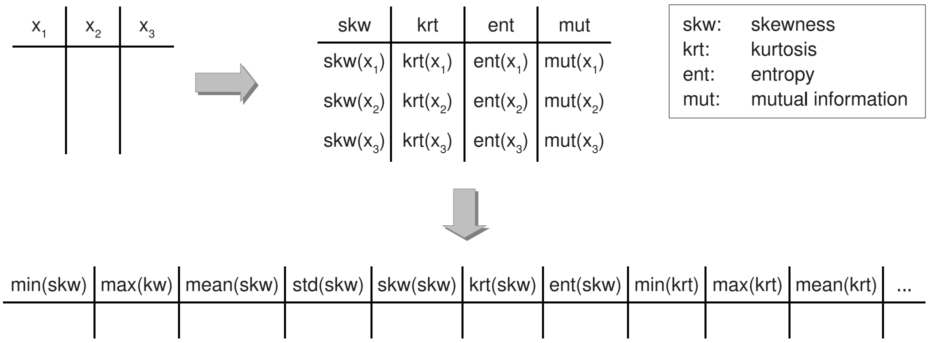
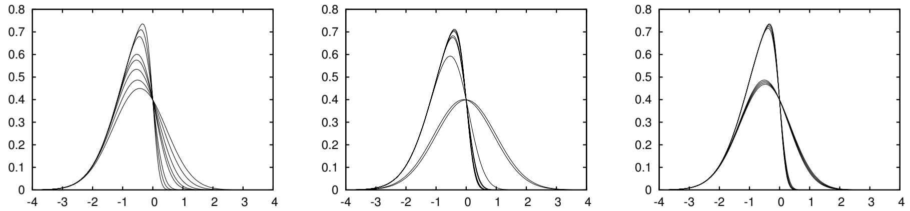

x_1 x_2 x_3
skw krt ent mut
skw(x_1) krt(x_1) ent(x_1) mut(x_1)
skw(x_2) krt(x_2) ent(x_2) mut(x_2)
skw(x_3) krt(x_3) ent(x_3) mut(x_3)
skw: skewness
krt: kurtosis
ent: entropy
mut: mutual information
min(skw) max(kw) mean(skw) std(skw) skw(skw) krt(skw) ent(skw) min(krt) max(krt) mean(krt) ...

Fig. 1. The presented approach uses two steps: first, the meta-features of each feature construct an intermediate dataset from which the final meta^2-features are calculated.



(a) mean skewness: 1.37
skewness of skewness: -0.18
kurtosis of skewness: -1.29

(b) mean skewness: 1.38
skewness of skewness: -1.00
kurtosis of skewness: -0.84

(c) mean skewness: 1.37
skewness of skewness: -0.00
kurtosis of skewness: -1.99

Fig. 2. The distribution of eight features for three artificial datasets: While the mean skewness is almost the same, the meta^2-features show significant differences.

both the skewness of the skewness values and the kurtosis of the skewness values show a significant difference.

Since the approach leads to an increased amount of meta-features while the usefulness of each single meta-feature is not proven, an automatic feature selection method should be applied in order to select the most useful ones. It was previously shown that automatic feature selection can improve the performance of meta-learning [11, 14].

3 Conclusion

We presented a novel approach of constructing more informative meta-features using a two-stage method based on traditional meta-features. The proposed meta^2-features are able to describe differences over datasets that are not accessible using the typically used mean of meta-measures, only. An additional feature selection method is suggested in order to automatically select the most useful measures.

References

1. Ali, S., Smith, K.A.: On learning algorithm selection for classification. Applied Soft Computing 6, 119–138 (January 2006)
2. Bensusan, H., Giraud-Carrier, C.: Discovering task neighbourhoods through landmark learning performances. In: Proc. of the 4th European Conf. on Principles of Data Mining and Knowledge Discovery. pp. 325–330 (2000)
3. Bensusan, H., Giraud-Carrier, C., Kennedy, C.: A higher-order approach to meta-learning. In: Proc. of the ECML'2000 workshop on Meta-Learning: Building Automatic Advice Strategies for Model Selection and Method Combination. pp. 109–117 (June 2000)
4. Brazdil, P.B., Soares, C.: Zoomed ranking: Selection of classification algorithms based on relevant performance information. In: Proc. of Principles of Data Mining and Knowledge Discovery PKDD. pp. 126–135 (2000)
5. Castiello, C., Castellano, G., Fanelli, A.M.: Meta-data: Characterization of input features for meta-learning. In: Torra, V., Narukawa, Y., Miyamoto, S. (eds.) Modeling Decisions for Artificial Intelligence, Lecture Notes in Computer Science, vol. 3558, pp. 295–304 (2005)
6. Engels, R., Theusinger, C.: Using a data metric for preprocessing advice for data mining applications. In: Proc. of the European Conf. on Artificial Intelligence. pp. 430–434 (1998)
7. Michie, D., Spiegelhalter, D.J., Taylor, C.C.: Machine Learning, Neural and Statistical Classification. Ellis Horwood (1994)
8. Peng, Y., Flach, P., Soares, C., Brazdil, P.: Improved dataset characterisation for meta-learning. In: Lange, S., Satoh, K., Smith, C. (eds.) Discovery Science, Lecture Notes in Computer Science, vol. 2534, pp. 193–208 (2002)
9. Pfahringer, B., Bensusan, H., Giraud-Carrier, C.: Meta-learning by landmarking various learning algorithms. In: Proc. of the 17th Int. Conf. on Machine Learning. pp. 743–750 (2000)
10. Reif, M., Shafait, F., Dengel, A.: Meta-learning for evolutionary parameter optimization of classifiers. Machine Learning 87(3), 357–380 (2012)
11. Reif, M., Shafait, F., Goldstein, M., Breuel, T., Dengel, A.: Automatic classifier selection for non-experts. Pattern Analysis and Applications (2012), 10.1007/s10044-012-0280-z
12. Segrera, S., Pinho, J., Moreno, M.: Information-theoretic measures for meta-learning. In: Corchado, E., Abraham, A., Pedrycz, W. (eds.) Hybrid Artificial Intelligence Systems, Lecture Notes in Computer Science, vol. 5271, pp. 458–465 (2008)
13. Spiliopoulou, M., Kalousis, A., Faulstich, L.C., Theoharis: Noemon: An intelligent assistant for classifier selection. In: 13th German Workshop on Machine Learning (August 1998)
14. Todorovski, L., Brazdil, P., Soares, C.: Report on the experiments with feature selection in meta-level learning. In: Brazdil, P., Jorge, A. (eds.) Proceedings of the PKDD-00 Workshop on Data Mining, Decision Support, Meta-Learning and ILP: Forum for Practical Problem Presentation and Prospective Solutions. pp. 27–39 (September 2000)