Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 25–28
DOI: 10.15439/2018F87
ISSN 2300-5963 ACSIS, Vol. 15

Data Compression Measures for Meta-Learning Systems.

Marcin Blachnik
Silesian University of Technology
Department of Applied Informatics
Katowice, ul. Krasińskiego 8, Poland
Email: marcin.blachnik@polsl.pl

Mirosław Kordos
University of Bielsko-Biala
Department of Computer Science and Automatics
Bielsko-Biała, ul. Willowa 2, Poland
Email: mkordos@ath.bielsko.pl

Sławomir Golak
Silesian University of Technology
Department of Applied Informatics
Katowice, ul. Krasińskiego 8, Poland
Email: slawomir.golak@polsl.pl

***Abstract*—An important issue in predictive modeling is model selection. This process is time consuming and can be simplified with meta-learning. However, meta-learning systems need appropriate data descriptors for proper functioning. One of them are data compression measures which can be extracted out of the instance selection methods. When we only need to estimate the classification accuracy of the model, the compression obtained from instance selection is a good approximator, but when we need to estimate other performance measures such as the precision and sensitivity then the quality of the estimated performance drops. To overcome this issue we propose a new type of compression measure: the *balanced compression* which is sensitive to the class label distribution and shows high correlation with precision and sensitivity of the final classifiers. We also show that the application of the *balanced compression* as a meta-learning descriptor allows for precise assessment of the model performance, as proved by the presented experimental evaluation.**

I. Introduction

NOWADAYS, meta-learning [1], [2] is gaining more and more popularity. It is aimed at speeding up the prediction model construction which consists of model selection and model parameters optimization. The model selection process can be done without actually training the given model, by using other meta-model which assesses the quality of the data and estimates the performance of the desired classier or returns a ranking.

As shown in [3], [4], a good indicator that characterizes the dataset quality is the compression of the dataset obtained with instance selection algorithms [7]. It is defined as: $Cmp = 1 - \frac{\|\mathbf{P}\|}{\|\mathbf{T}\|}$ where $\mathbf{T} = [\{\mathbf{x}_1, y_1\}, \{\mathbf{x}_2, y_2\}, \ldots \{\mathbf{x}_n, y_n\}]$ is a training dataset that consists of n training instances $\{\mathbf{x}, y\}$, where $\mathbf{x} \in \Re^m$ and y is a label which takes one of l symbols, and the dataset $\mathbf{P}$ is a subset of the instances from $\mathbf{T}$ selected by the instance selection algorithm, so that $\mathbf{P} \subset \mathbf{T}$.

The main idea of using compression as a meta-learning descriptor (also called meta-attribute) is based on the observation that a dataset in which there is a lot of regularity can be compressed well, and thus high prediction accuracy should be achievable, while a dataset containing a lot of irregularities and a lot of noise will have a low compression ratio. Moreover, the instance selection methods are often used at the stage of data preprocessing, which means that the value of compression is obtained without additional computational cost. Some algorithms, including *CNN* and *ENN*, have been identified as the most useful for predicting the final model performance [4]. For example, the correlation between the compression ratio and the accuracy of the *k*NN, Gaussian SVM and Random Forest, obtained for *CNN* and *ENN* instance selection methods is above 0.9. However, the research carried out so far has focused only on the classical definition of the measure of prediction accuracy expressed as the ratio of the correctly classified examples to all evaluated examples.

It turns out that although correlation between compression and classification accuracy is very high, the correlation between compression and other measures of classifier performance is much weaker. We refer to such measures as the average precision (also known as the balanced accuracy), or the average sensitivity also called recall, which are especially important in the context of unbalanced classification problems.

This work addresses this problem by introducing a new type of compression, so-called balanced compression, which takes into account the number of rejected instances which belong to particular classes. The balanced compression linearizes the relationship between compression and precision and between compression and sensitivity. Implementing these measures allows to enhance the meta-learning system performance.

II. Instance Selection Algorithms

As it was mentioned, the compression achieved by instance selection can be used as a measure of the dataset quality. We presented also an intuitive dependence, which indicates that stronger compression is connected with greater regularity of the decision boundaries in the dataset, and at the same time it is easier for the classifier to reconstruct the desired decision boundary. In practice, however, this depends on the particular instance selection algorithm. These algorithms can be divided into three basic groups: condensing methods, noise filters and hybrid methods. Condensing methods are a set of algorithms used to reduce the dataset size, where the only criterion is maximization of compression while maintaining comparable prediction accuracy. A typical example is the *CNN* algorithm [8]. *CNN* was developed for use with the *k*NN classifier to reduce the computational complexity. The acceleration is accomplished by eliminating (compressing) unnecessary instances in the dataset. However, this does not help increasing the prediction quality of this classifier. There are algorithms that allow for stronger compression at the same accuracy level, e.g. evolutionary based instance selection [6]. However, as the evolutionary approach belongs to the hybrid group, the correlation we observed although is still significant, is weaker then that obtained with *CNN*.

Noise filters, on the other hand, are a set of algorithms created with the purpose of finding and removing training instances that constitute noise in a dataset. An example and historically the first noise filter is the *ENN* algorithm designed to improve prediction accuracy of 1-NN [9]. *ENN* was also developed to work with the *k*NN algorithm. Its operation is based on the analysis of the closest neighborhood of a given instance and checking if the nearest neighbors will vote for the examined instance in accordance with its label. If not, then the instance is removed.

Also generalizations of these algorithms were proposed, where different classifiers, not only *k*NN can be embedded into the instance selection process [10]. However, in the experiments presented in this paper only instance selection based on 1-NN will be considered.

The third group of instance selection algorithms are hybrid methods. They combine the properties of the first two groups. They start by filtering out the noisy samples from the data and then condense the remaining dataset.

As it was shown in [4], each group of instance selection methods behaves differently with regard to the prediction accuracy. For the condensing methods, an increase in compression corresponds to an increase in prediction accuracy. In the case of noise filter methods, this relation is reversed, because the noise filters regularize and clean the datasets from noise. Thus more removed instances indicate here more noisy dataset, which means that with the increase of compression of the noise filters, the reduction of prediction accuracy is observed. The last group - hybrid methods combine both elements. This causes that the relationship between compression and prediction accuracy gets much weaker or totally disappear, because the properties of condensation methods are canceled out by the properties of noise filters. This causes that only the instance selection methods, which obtain different compression depending on noise in data find application in estimating the prediction accuracy.

III. Balanced Compression measure

As mentioned in the introduction, for unbalanced classification problems usually classical accuracy measure is not used and rather other performance measures are evaluated like average precision or average sensitivity. The purpose of these measures is to reflect the quality of the prediction model in the context of the number of instances in individual classes. A similar situation occurs in the case of compression measures. The commonly used compression measure ignores the number of rejected instanced within individual classes. It simply represents the ratio of the number of rejected samples to the size of the training set $\mathbf{T}$, thus, this measure is similar to the classical accuracy used in prediction systems. The natural conclusion from this is that we should adapt the measure of compression to data with unbalanced class distribution, so that the measure not only indicates the number of rejected samples but also the number of rejected samples within individual classes. It can bring tangible benefits in the form of additional information about the nature of the classification problem, in particular in the context of meta-learning systems.

An important difference between accuracy and compression is the fact, that in contrast to the evaluation of the accuracy of the classifier, in the case of compression we do not have the confusion matrix and the values resulting from it like *False Positives* or *False Negatives*. It is because instance selection methods do not perform prediction, instead we only have information which instances were selected and which rejected, so we do not know what type of error occurred. Therefore, the only factor possible to determine is the level of class c_i compression defined as $\frac{\|y_{\mathbf{T}}==c_i\|-\|y_{\mathbf{P}}==c_i\|}{\|y_{\mathbf{T}}==c_i\|}$, where $\|y_{\mathbf{T}} == c_i\|$ denotes the number of samples in the training set $\mathbf{T}$ which belong to class c_i and $\|y_{\mathbf{P}} == c_i\|$ denotes the number of instances in the dataset $\mathbf{P}$ (after instance selection) which belong to class c_i.

Based on this class compressions we define balanced compression as an average over all classes

$$Cmp_{Bal} = \frac{1}{l}\sum_{i=1}^{l}\frac{\|y_{\mathbf{T}} == c_i\| - \|y_{\mathbf{P}} == c_i\|}{\|y_{\mathbf{T}} == c_i\|} \tag{1}$$

where l denotes the number of classes. This measure can be also generalized by introducing class weights denoted as w_i which describes importance of particular class, so the balanced compression takes the form:

$$Cmp_{Bal} = \frac{1}{\sum w_i}\sum_{i=1}^{l} w_i\frac{\|y_{\mathbf{T}} == c_i\| - \|y_{\mathbf{P}} == c_i\|}{\|y_{\mathbf{T}} == c_i\|} \tag{2}$$

In the conducted experiments we assumed equal values of the weights $\underset{i=1\ldots l}{\forall} w_i = 1$.

IV. Experiments and Results

In order to verify the usefulness of the proposed *balanced compression* in the context of meta-learning systems, we carried out an experimental evaluation on 45 datasets obtained from Keel Project [11] using three popular classifiers: *k*NN, linear SVM and Random Forest. The experiments were performed with RapidMiner and the Information Selection package developed by the authors of this paper, which is available from the RapidMiner Marketplace and on the website www.prules.org [12]. The experiments were divided into two parts. In the first part the correlation measure was evaluated between compression measures and performance measures of the evaluated classifiers. It indicates how the new compression measure reflects the obtained classification performances. In the second part a real meta-learning system was constructed which is designed to predict performance of the three classifiers. The meta-learning system utilizes meta-attributes which are based on compressions obtained by both *CNN* and *ENN*.

A. Relationships Between Compression and Various Performance Measures

The first part of the experiments consists of two stages. In stage I, the three performance measures (accuracy, average precision and average sensitivity) for each of the 45 datasets were estimated using the cross-validation procedure. This stage also included parameter optimization for all evaluated classifiers (k for *k*NN, C- for linear SVM and the number of trees for Random Forest). In stage II, each dataset was compressed using the two previously described algorithms *ENN* and *CNN*, each time measuring both compression and balanced compression. The obtained results were then used to calculate Pearson's correlation coefficient between given type of compression and the type of classification performance measure independently for each classifier. Obtained correlations were collected in Tab. I for *CNN* instance selection, and in Tab. II for *ENN* instance selection algorithm.

Table I: Correlation between two types of compression obtained for *CNN* and the three performance measures for *k*NN, Linear-SVM and Random Forest

Compression type	Performance measure	*k*NN	SVM	Random Forest
Compression	Accuracy	0.937	0.902	0.900
Compression	Precision	0.783	0.662	0.767
Compression	Recall	0.738	0.640	0.767
Balanced compression	Precision	0.920	0.794	0.880
Balanced compression	Recall	0.932	0.808	0.880

Table II: Correlation between two types of compression obtained for *ENN* and the three performance measures for *k*NN, Linear-SVM and Random Forest

Compression type	Performance measure	*k*NN	SVM	Random Forest
Compression	Accuracy	-0.965	-0.924	-0.917
Compression	Precision	-0.774	-0.672	-0.745
Compression	Recall	-0.758	-0.661	-0.765
Balanced compression	Precision	-0.935	-0.844	-0.883
Balanced compression	Recall	-0.981	-0.863	-0.895

The results in the tables indicate that the correlation between classical compression and prediction accuracy is very high and ranges from 0.917 to 0.965 for the *ENN* algorithm and from 0.900 to 0.937 for the *CNN* (here for simplicity we evaluate absolute values of the correlation as the sign does not matter). However, changing the measure of the prediction quality to average precision or average sensitivity causes the correlation coefficient to drop rapidly to a level between 0.64 and 0.77 depending on the method of instance selection and on the classifier. Changing compression to the balanced compression results in a significant increase in the correlation coefficient, which for the *k*NN classifier again exceeds 0.9, and for the other classifiers varies between 0.8 and 0.88. This is a significant improvement over the correlation coefficients obtained with standard compression.

B. Meta-system - compression-based estimation of prediction quality

Meta-learning systems are used for the estimation of quality of predictive models [13], [1], [14], [15]. In these systems, for a known dataset repository, which consists of n_r datasets, the prediction performance of the selected classifier is estimated and the meta-attributes describing the properties of each of these datasets are extracted [16]. Next, a meta-set is created. The meta-set consists of the extracted meta-attributes (an input vector of the meta-learning system) and labels that express the accuracy of the given model, for which we would like to estimate the accuracy. Therefore, the meta-set consists of n_r samples, where a single instance describes one dataset from the repository. So we obtain a typical regression problem, because labels in the meta-set represent numerical values (accuracies). In the next step, the meta-set is used to build a meta-model, a model capable of estimating prediction accuracy for a given, previously unknown data set. When applying a meta-model to a new data set, it is necessary in the first step to determine the meta-attributes, create a record from them, and then pass them to the meta-model input. The meta-model then returns the estimated accuracy. Another commonly used solution is learning the meta-ranking model, where the meta-model returns the ranking of the best models or just the best prediction model [17].

It was shown in [4] that the use of compressions as meta-attributes lead to an improvement in the quality of the estimated accuracy in comparison to the classic meta-attributes used in the MLWizzard system [15]. Therefore, in the experiments a meta-system based only on the data set compression measures is constructed.

As a meta-model, Generalized Linear Model was used. In total we had 9 meta-models (for each of the three performance measures and for each of the tree classifiers). The meta-model was tested using the 5x10 cross-validatin procedure. The quality of the whole system was evaluated using *RMSE* calculated between predicted and real prediction performance. The obtained results are presented in Tab. III.

The results are placed in two columns. The first column contains the results obtained using classical compression of both *CNN* and *ENN* as meta-attributes, and the second column contains the results obtained with a balanced compression. For each of the tested classifiers, the three measures of accuracy (accuracy, average sensitivity and average precision) were estimated, and Welch's t-test [18] was used to determine if the results are statistically significantly different at $\alpha = 0.05$. The symbol (+) indicates results which are significantly better.

The obtained results clearly indicate that for meta-learning systems where the task is to estimate classical accuracy, the standard compression measure gives better results. However, when the aim of the process is to estimate average precision or average sensitivity, a much better solution is to use the balanced compression; each time the results obtained using balanced compression were statistically significantly better than those obtained with standard compression.

Table III: Results of the meta-learning system. The columns represent RMSE of the meta-model aimed at estimating classification performance of three classifiers: *k*NN, Linear SVM, and Random Forest using two meta-sets which consisted of: clasical compression based meta-attributes (column 1) and balanced compression - based meta-attributes (column 2)

		Compression RMSE±std	Balanced Compression RMSE±std
*k*NN	Accuracy	0.0328±0.0207(+)	0.0799±0.0400
	Recall	0.1336±0.0529	0.0703±0.0575(+)
	Precision	0.1265±0.0607	0.0884±0.0628(+)
SVM	Accuracy	0.0640±0.0296(+)	0.0910±0.0416
	Recall	0.1523±0.0610	0.1130±0.0600(+)
	Precision	0.1453±0.0711	0.1171±0.0622(+)
Random Forest	Accuracy	0.0437±0.0281(+)	0.0739±0.0408
	Recall	0.1276±0.0632	0.0954±0.0711(+)
	Precision	0.1284±0.0644	0.0979±0.0711(+)

The prediction quality of the *k*NN model can be estimated more precisely than those of SVM or Random Forest, which is natural, as the instance selection methods internally use the nearest neighbor mechanism to evaluate each of the instances. Random Forest ranked lower than *k*NN in terms of performance estimation but ranked higher than SVM. SVM's high performance estimation error was caused by the fact that the SVM considered in this study utilized a linear kernel, and thus it was a linear classifier, while Random Forest is a nonlinear classifier. By their very nature, methods of instance selection are nonlinear, and thus they can overestimate the results obtained by the linear model.

V. Conclusions

In this study we have shown that compression forms a strong linear relationship with the standard prediction accuracy. We have also shown that other measures of prediction quality do not correlate strongly with the standard compression obtained by instance selection.

To address this problem, we proposed a modified measure of compression called balanced compression. The purpose of balanced compression is to express the characteristics of the dataset preserving distribution of the class labels. This allowed to obtain almost linear relationship between the balanced compression and the accuracy measures such as average precision and average sensitivity. The importance of this linear relationship can be efficiently used in meta-learning systems, where the balanced compression allowed for a significant improvement in the estimation of average precision and average sensitivity compared to estimation performed using standard compression.

Acknowledgment

This research was supported by the Silesian University of Technology, Grant 11/040/RGJ17/0014.

References

[1] N. Jankowski, W. Duch, K. Grąbczewski, *Meta-learning in computational intelligence*. Springer Science & Business Media, vol. 358, 2011.

[2] L. Kotthoff, Ch. Thornton, H. Hoos, F. Hutter,K. Leyton-Brown, "Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA," *JMLR*, vol. 18, no. 1, pp. 826–830, 2017.

[3] M. Blachnik, "On the relation between knn accuracy and dataset compression level," *LNAI*, vol. 9692, pp. 541–551, 2016.

[4] M. Blachnik, "Instance selection for classifier performance estimation in meta learning," *Entropy*, vol. 19, no. 11, p. 583, 2017.

[5] M. Kordos, M. Blachnik, J. Kozłowski, M. Perzyk, O. Bystrzycki, M. Gródek, A. Byrdziak, Z. Motyka, "A Hybrid System with Regression Trees in Steelmaking Process," *LNAI*, vol. 6678, pp. 222-229, June 2011.

[6] M. Kordos, "Optimization of Evolutionary Instance Selection," *LNAI*, vol. 10245, pp. 359-369, ICAISC, June 2017

[7] S. García, J. Derrac, J. R. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: Taxonomy and empirical study," *IEEE Trans Pattern Anal and Mach Intell*, vol. 34, no. 3, pp. 417–435, 2012.

[8] P. Hart, "The condensed nearest neighbor rule." *IEEE Trans. on Information Theory*, vol. 16, pp. 515–516, 1968.

[9] D. Wilson, "Assymptotic properties of nearest neighbour rules using edited data." *IEEE Trans. on Systems, Man, and Cybernetics*, vol. SMC-2, pp. 408–421, 1972.

[10] M. Kordos, M. Blachnik, and S. Białka, "Instance selection in logical rule extraction for regression problems," *LNAI*, vol. 7895, pp. 167–175, 2013.

[11] F. Herrera, "Keel, knowledge extraction based on evolutionary learning," http://www.keel.es, 2005, spanish National Projects TIC2002-04036-C05, TIN2005-08386-C05 and TIN2008-06681-C06. [Online]. Available: http://www.keel.es

[12] M. Blachnik and M. Kordos, "Information selection and data compression rapidminer library," in *Machine Intelligence and Big Data in Industry*. Springer, 2016, pp. 135–145.

[13] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc., 2015, pp. 2962–2970.

[14] M. Kozielski, "A meta-learning approach to methane concentration value prediction," in *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery*. Springer, 2015, pp. 716–726.

[15] M. Reif, F. Shafait, and A. Dengel, "Meta-learning for evolutionary parameter optimization of classifiers," *Machine Learning*, vol. 87, no. 3, pp. 357–380, 2012.

[16] F. Pinto, C. Soares, and J. Mendes-Moreira, "Towards automatic generation of metafeatures," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2016, pp. 215–226.

[17] Q. Sun and B. Pfahringer, "Pairwise meta-rules for better meta-learning-based algorithm ranking," *Machine learning*, vol. 93, no. 1, pp. 141–161, 2013.

[18] B. L. Welch, "The generalization ofstudent's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.