Information-Theoretic Measures for Meta-learning

Saddys Segrera, Joel Pinho, and María N. Moreno

Department of Computer Science and Automatics, University of Salamanca
Plaza de la Merced s/n, 37008, Salamanca, Spain
{saddys,joelpl,mmg}@usal.es

Abstract. Information-theoretic measures are suitable to characterize datasets with discrete attributes (or continuous which can be transformed). They can find information that can be decisive in order to analyze the behavior of different learning algorithms with specific datasets. The objective of the work presented in this paper is to study by means of three similar datasets from UCI Repository Machine Learning, the possible reasons for which breast-cancer-wisconsin dataset, in comparison with other 20 datasets, showed in a previous research that Stacking by Meta-Decision Trees (MDT) was significant better than all other multiclassifier models, including Stacking by Multi-Response Linear Regression (MLR). In our experiments the proportion of missing values, among other significant changes in different measure values, provided evidences about the possible origin of the different behaviors presented by these multiclassifier schemes depending on data characteristics.

Keywords: meta-learning, classification, stacking.

1 Introduction

Meta-learning is defined as the process of supervised learning that takes place from the information generated by the initial or base classifiers. This means, a technique to unify the results of multiple classifiers. The idea is to generate a system that performs the base classifiers functionality and increase the precision by means of the improvement of the form in which they correlate themselves [1], which leads to an efficient reduction of the space of incorrect predictions [2].

One technique addressed to study meta-learning is the characterization of datasets. Suitable data characterization is very important for the meta-learning. The complexity of data mining tasks is related to the characteristics of datasets and the inductive bias of learning algorithms [3].

The objective of this paper is the characterization of three datasets. Breast-cancer-wisconsin dataset was the only of 21 datasets from UCI Machine Learning Repository in the schemes (multiclassifiers) used by [4] in what Stacking [5] with meta-decision trees (MDT) was significantly better in all the cases. Inspired by this results, we compared breast-cancer-wisconsin with other two datasets (hepatitis and vote), which have similar composition of attributes, but had different behaviors in the research mentioned before.

E. Corchado, A. Abraham, and W. Pedrycz (Eds.): HAIS 2008, LNAI 5271, pp. 458–465, 2008.

We want to search similarities and differences in behaviors of Stacking with MDT and Stacking with Multi-response linear regression (MLR) and to provide some relative contributions related to information-theoretic measures for Stacking meta-learning.

This paper has been organized as following. In section 2 we summarize a group of related works based in dataset characterization techniques. The information-theoretic measures used in our experiments to the dataset characterization are described in section 3. The characterization of breast-cancer-wisconsin, hepatitis and vote datasets from UCI Machine Learning Repository and the analysis of the experimental results obtained are included in section 4. Finally, the conclusions are presented in section 5.

2 Related Works

Dataset characterization techniques are used in order to describe the problem that will be studied, including simple measures as number of attributes and number of classes; statistical measures as variance and average of numerical attributes; and measures based in theory of information as entropy of classes and attributes [6]. Nevertheless, there exists the need to improve the efficiency of the meta-learning developing better meta-attributes and selecting those which provide more information.

The first initiative to characterize datasets to predict the execution of a classification algorithm was made in [7]. Since that, three main strategies have developed for dataset characterization [8], and to suggest in this way what algorithm is most adapted for a specific dataset. One of them is the technique that describes the characteristics of the dataset using statistical and informative measures [9, 10]. STATLOG project [10] allowed the description of datasets by their information and statistical properties. The authors identified three categories of data characteristics: simple, statistical and information theory based measures. METAL [11] project is another example that allows characterizing data for meta-learning by means of different measures. Statistical characteristics are mainly appropriated for continuous attributes, while information-theoretic measures are more appropriated for discrete attributes [3]. On the other hand, in [3, 12] a second strategy is used for dataset characterization. In this case, the characteristics of the induced hypothesis as a way to represent the own dataset are considered. The third strategy consists of characterizing a dataset using the behavior of a system of simplified classifiers named landmarking [13].

Whereas other approaches typically describe a data set with statistical measures and information of attributes, landmarking proposes to enrich such description with fast and easy operation measures from simple learning algorithms. Learning algorithm profiles have been also used in meta-learning. These profiles consist of metalevel feature-value vectors which describe learning algorithms from the point of view of their representation and functionality, efficiency, robustness, and practicality. For certain characteristics related to functionality (attribute types, cost handling), algorithm specifications are given by expert users. Characteristics related to efficiency (learning and classification time and space) and robustness (scalability, resistance to missing values, noise, irrelevant and redundant attributes) can only be extrapolated from multiple executions of these algorithms over a wide variety of datasets [14].

3 Information-Theoretic Measures

Various papers [15, 16] have introduced the use of measures to characterize the data complexity and to relate such descriptions to classifier performance for two classes.

Information-theoretic measures are suitable to characterize discrete attributes. We used in our experiment: the entropy of the class label (ClassEnt), the entropy of all attributes (AttrEnt), the mutual information (entropy) of class and attributes (MutualInf), the joint entropy (JointEnt), the equivalent number of attributes (EquivAttr), proportion of the equivalent number of attributes (PropEquivAttr), the noise signal ratio (NoiseSR), the proportion of missing values (PropMV), proportion of number of examples with missing values (PropExMV) and a statistical measure that is the standard deviation of classes (StdDClass).

Let X be a random variable taking values x in X with distribution $p(x)=Pr[X=x]$. The *entropy H(X)* of a random variable X (the label of the problem) is defined by:

$$H(X) = -\sum_{x \in X} p(x) \ \log_b p(x) \tag{1}$$

This is also denoted by $H(p)$ and measures the average uncertainty of a random variable X, b is the base of the logarithm used. Possible values of b are 2, e, and 10. The unit of the information entropy is bit for b=2. Then, in this case the entropy of the class label values belong to the interval $[0, \log_2 n]$, being n the number of the different values of the label. It means the maximum value of entropy of the class label for these three datasets is 1 since n=2.

The entropy of all attributes and the label (*Joint entropy*) measures the total entropy. It is the sum of the individual entropy of all variables appearing in the dataset.

The entropy of a collection of attributes is an average of the entropy over all the attributes, which is taken as a global measure of entropy of the attributes collectively [10].

Mutual information expresses the mutual dependency of the attributes and the label. It is the amount of information that can be obtained about the label by observing the attributes. The mutual information between X and Y is defined by:

$$MutualInf(X,Y) = \sum_{x \in X} p(x) \ \sum_{y \in Y} p(y|x) \log \frac{p(y|x)}{p(y)} \tag{2}$$

The *equivalent number of attributes* estimates the number of attributes that are needed to determine the value of the label variable. If the number of relevant attributes that are provided by the dataset is larger than the value of this measure, there exists a good chance to learn a good classification algorithm [17]. The expression is:

$$EquivAttr = \frac{H(X)}{MutualInf(X,Y)} \tag{3}$$

The *proportion of the equivalent number of attributes* is calculated by the equivalent number of attributes divided by the number of attributes excluding the label.

Noise signal ratio is the amount of irrelevant information; large values of the ratio indicate that the dataset contains a large amount of irrelevant information that may be reduced [18]. It can be calculated by:

$$NoiseSR = \frac{\overline{H(X)} - \overline{MutualInf(X,Y)}}{\overline{MutualInf(X,Y)}} \tag{4}$$

The *proportion of missing values* is the ratio of the total missing values in the dataset between the number of all values in it.

The *proportion of number of examples with missing values* is the division of the total examples with missing values in the dataset between the number of examples.

4 Experimental Results

In this section the characteristics of breast-cancer-wisconsin (BCW) dataset are studied, because it is the only dataset from UCI Machine Learning Repository used in [4] where Stacking by MDT was significantly better than the other schemes used, even better than Stacking by MLR. We want to know the properties of this dataset in order to identify when Stacking by MDT is the best choice rather than Stacking by MLR. The objective is to search and to contribute with new elements in the Stacking meta-learning. We want also to find features that can be employed in another similar technique in order to improve meta-learning.

This dataset has 699 examples and 11 attributes (10 plus the label). There are 16 missing attribute values. The label for classification is composed by diagnosis classes. It has 2 classes (benign and malignant). The other attributes belong to the integer number set but they can be easily discretized because all of them take values in the interval [1,10], then we made the transformation. The names of them are: sample code number (it is not considered), clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. The class distribution is 458 examples (65.5%) of the "benign" class and 241 examples (34.5%) of the "malignant" class. We used Stacking with three base classifiers in our experiments: decision tree, nearest neighbor and Naïve Bayes. Furthermore, MLR was chosen to learn at the meta-level in one case and MJ4.8 in another.

The dimensionality of the BCW input space was reduced due to the removing of marginal adhesion, single epithelial cell size and bare nuclei attributes because there is a multivariate dependency among them. This strategy was also used in [19] to characterize the same dataset for the rule extraction. The missing values disappear when the reduction of the number of attributes is done.

We used the 10 measures described in the previous section and their values were obtained by METAL-MLEE (Machine Learning Experimentation Environment) [20], which is a software package developed for the METAL Project that helps in obtaining the meta-data for new datasets and different algorithms in performing meta-learning.

Two other datasets (hepatitis and vote) from UCI Machine Learning Repository for task classification were selected with only two class label, missing values and similar number of continuous attributes, as BCW dataset, in order to make comparisons.

Hepatitis dataset has continuous and discrete attributes. Vote dataset only has discrete attributes.

Hepatitis dataset did not show the same results than BCW dataset in [4]. Stacking by MDT for hepatitis dataset did not present significant difference in relation to the other seven schemes studied in that research. However, Stacking by MDT for vote dataset was significant better than the other schemes, excluding Stacking by MLR in the same experiment. Significant differences were not found between Stacking by MDT and Stacking by MLR for this dataset.

All datasets achieved high entropy of class label, although vote and BCW (in this order) highlighted because they reached values near to 1. The difference between them was not significant (0.033), see Table 1.

The highest value of the joint entropy was reached by BCW (over 2), while hepatitis and vote datasets obtained similar values.

The entropy of all attributes in BCW dataset is superior to the others. It is more than the double of the values of this metric obtained for hepatitis and vote datasets. It means the attributes of BCW dataset provide as average more information than the others.

In Table 1, we can also observe BCW dataset gets the best value of the useful information that each attribute has about the class label (mutual information).

The proportion of the equivalent number of attributes had a similar behavior in BCW with 9 attributes and vote, being the value for hepatitis dataset almost the double of the other datasets. When we executed the reduction to 6 attributes in BCW, this dataset obtained the highest value (45.3%) and hepatitis dataset got a value very close to 40%.

Table 1. Data characteristics of BCW and other two datasets using METAL-LEE

Measure	BCW with 9 attributes	BCW with 6 attributes	Hepatitis	Vote
ClassEnt	0.929	0.929	0.735	0.962
AttrEnt	2.251	2.299	1.052	0.990
MutualInf	0.512	0.507	0.103	0.304
JointEnt	2.668	2.721	1.683	1.649
EquivAttr	1.813	1.832	7.142	3.165
PropEquivAttr	0.201	0.453	0.376	0.198
NoiseSR	3.392	3.533	9.225	2.257
PropMV	0.002	0	0.059	0.053
PropExMV	0.023	0	0.484	0.467
StdDClass	0.155	0.155	0.293	0.114

Hepatitis dataset attained the greatest value of the noise signal ratio. There is a significant difference with the other two datasets: 6.968 with regard to vote and 5.692 in relation to BCW for 6 attributes.

If we analyze the results of BCW with 9 attributes, it is possible to see in Table 1 that the less percentage of missing values regarding to all values in the dataset was reached by BCW with 0.2%, it is very low, while hepatitis and vote datasets have values close to 6% and 5%, respectively. There is a very significant difference for the proportion of number of examples with missing values between BCW and the other

datasets. BCW only has a 2.3% of examples with missing values and the rest has almost the 50%.

However, BCW with 6 attributes has not missing values, and then all the measures in relation to missing values got values 0.

After the analysis of the measure calculations it is possible to infer there is most probability datasets with few or without missing values achieve better results in classification with Stacking by MDT than Stacking by MLR. Together with this, it is necessary the entropy of the label is very close to the maximum value, the average of the entropy of all attributes, the joint entropy and mutual information obtain high value, and the equivalent number of attributes and noise signal ratio get very low values.

We wanted also to know the ranking of the three algorithms used as base classifiers to predict the label of BCW dataset and therefore, to obtain the meta-attributes.

WekaMetal [21] is a meta-learning extension to the data mining package Weka [22]. It has been used in order to obtain the rankings of IBk (nearest neighbor), decision tree (J48) and Naïve Bayes classifiers for BCW dataset. The ranker selected for this study was Adjusted Ratio of Ratios (ARR), based on expected accuracy and time performance; see the results in Table 2.

Table 2. Ranking of three algorithms for BCW dataset using WekaMetal

Ranking	Algorithm	Measure multi-criteria
1	Naïve Bayes	2.6078
2	J48	0.6659
3	IBk	0.6339

In our experiments, Stacking by MDT and Stacking by MLR had the same accuracy value using BCW with Weka (Fig. 1), and the use of computational resources of the first multiclassifier is almost two times that the second one. In all experiments we carried out ten-fold cross-validation. In Fig. 1, it is possible to show Naïve Bayes is, of course, faster than the two Stacking schemes, and the Naïve Bayes accuracy value

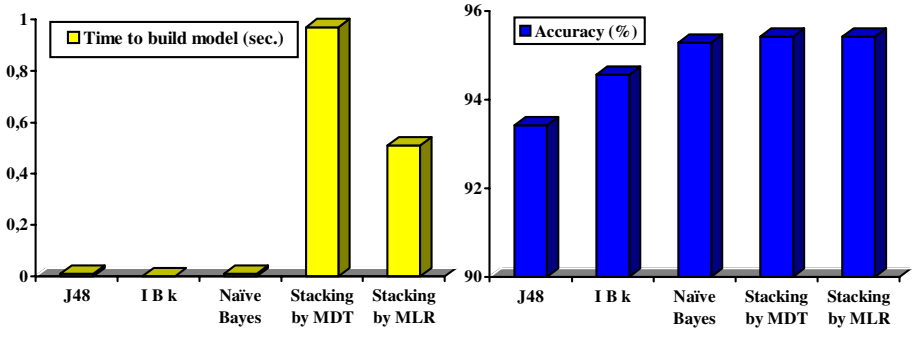

Fig. 1. Times to build models and classification accuracies for BCW dataset

is not significantly lower than the multiclassifiers. Then, the use of Stacking schemes could not be justified because there is a base classifier which is faster and obtains a similar accuracy value than the Stacking schemes used.

5 Conclusions

In our experiment, datasets with discrete attributes exclusively, in classification by Stacking MDT achieved similar behavior when compared to Stacking by MLR.

According the results obtained, it is possible to point out Stacking by MDT and Stacking by MLR have similar classification accuracy in datasets with 2 label values, when the entropy of the label is very close to the maximum value, the average of the entropy of all attributes, the joint entropy and mutual information values are high. Moreover, the equivalent number of attributes must be very low and also the noise signal ratio.

Important information-theoretic measures in this experiment were those related to the missing values. Very low percentages of missing values and a small number of examples with missing values took place only in BCW dataset.

Finally, for future work we are planning a study with other metrics based on information theory and landmarking approach, which can also be applied to data characterization. On the other hand, the incorporation of other meta-attributes to the meta-level could also be useful.

Acknowledgments

This paper has been done in the framework of the research project SAD64A07, supported by the Spanish *Junta de Castilla y León*.

References

1. Fan, D., Chan, P., Stolfo, S.: A comparative evaluation of Combiner and Stacked Generalization. In: Proceedings of AAAI 1996 Workshop on Integrating Multiple Learned Models, pp. 40–46 (1996)
2. Chan, P., Stolfo, S.: On the accuracy of Meta-learning for Scalable Data Mining. Journal of Intelligent Information Systems 8(1), 5–28 (1997)
3. Peng, Y., Flach, P., Brazdil, P., Soares, C.: Decision Tree-Based Data Characterization for Meta-Learning. In: Proceedings of the Second International Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning (IDDM 2002), pp. 111–122. Helsinki University Printing House (2002)
4. Zenko, B., Todorovski, L., Dzeroski, S.: A comparison of stacking with meta decision trees to other combining methods. In: Proceedings of the Fourth International Multi-Conference Information Society, vol. A, pp. 144–147. Jozef Stefan Institute, Ljubljana (2001)
5. Wolpert, D.: Stacked generalization. Neural Networks 5, 241–259 (1992)

6. Peng, Y., Flach, P., Soares, C., Brazdil, P.: Improved Dataset Characterisation for Meta-learning. In: Lange, S., Satoh, K., Smith, C.H. (eds.) DS 2002. LNCS, vol. 2534, pp. 141–152. Springer, Heidelberg (2002)
7. Rendell, L., Seshu, R., Tcheng, D.: Layered Concept Learning and Dynamically Variable Bias Management. In: Proceedings of the 10th International Joint Conference on Artificial Intelligence, pp. 308–314 (1987)
8. Vilalta, R., Giraud-Carrier, C., Brazdil, P.: Meta-Learning: Concepts and Techniques. In: Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Springer, Heidelberg (2005)
9. Köpf, C., Taylor, C., Keller, J.: Meta-analysis: from data characterisation for meta-learning to meta-regression. In: Proceedings of the PKDD-2000 Workshop on Data Mining, Decision Support, Meta-Learning and ILP (2000)
10. Michie, D., Spiegelhalter, D., Taylor, C. (eds.): Machine Learning, Neural and Statistical Classification, volume: Artificial Intelligence. Ellis Horwood (1994)
11. Kalousis, A., Hilario, M.: Model selection via meta-learning: a comparative study. In: Proceedings of the 12th International IEEE Conference on Tools with AI. IEEE Press, Los Alamitos (2000)
12. Bensusan, H., Giraud-Carrier, C., Kennedy, C.: A higher-order approach to meta-learning. In: Proceedings of the ECMLS 2000 Workshop on Meta-Learning: Building Automatic Advice Strategies for Model Selection and Method Combination, pp. 109–117 (2000)
13. Pfahringer, B., Bensusan, H., Giraud-Carrier, C.: Tell me who can learn you and i can tell you who you are: Landmarking various learning algorithms. In: Proceedings of the 17th International Conference on Machine Learning, pp. 743–750 (2000)
14. Hilario, M., Kalousis, A.: Building Algorithm Profiles for Prior Model Selection in Knowledge Discovery Systems. In: Proceedings of the IEEE SMC 1999, International Conference on Systems, Man and Cybernetics, Tokyo (October 1999)
15. Ho, T.K., Basu, M.: Complexity measures of supervised classification problems. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 289–300 (2002)
16. Bernardo, E., Ho, T.K.: On classifier domain of competence. In: Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, pp. 136–139 (2004)
17. Studer, R., Staab, H.A., Mädche, A., Jetter, U.: Vorlesung Knowledge Discovery. Data Characterization Tool (DCT). Institute AIFB, Universität Karlsruhe (1999)
18. Köpf, C.: Meta-learning: Strategies, Implementations, and Evaluations for Algorithm Selection. Dissertations in Database and Information Systems-Infix, vol. 91. IOS Press, Amsterdam (2006)
19. Taha, I., Gosh, J.: Characterization of the Wisconsin breast cancer database using a hybrid symbolic-connectionist system. Technical Report UT-CVI-TR-97007. The Computer and Vision Research Center, University of Texas, Austin (1996)
20. Petrak, J.: The METAL Machine Learning Experimentation Environment V3.0 (METAL-MLEE) Manual - Version 3.0. Austrian Research Institute for Artificial Intelligence (October 2002), `http://www.metal-kdd.org/`
21. Farrand, J.: WekaMetal. University of Bristol (February 2002), `http://www.cs.bris.ac.uk/~farrand/wekametal`
22. University of Waikato. Weka (1999), `http://www.cs.waikato.ac.nz/~ml/weka/index.html`