ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Introducing diversity among the models of multi-label classification ensemble

Lena Chekina, Lior Rokach and Bracha Shapira

Ben-Gurion University of the Negev - Dept. of Information Systems Engineering
and Telekom Innovation Laboratories
Beer-Sheva, 84105 - Israel

Abstract. A number of ensemble algorithms for solving multi-label classification problems have been proposed in recent years. Diversity among the base learners is known to be important for constructing a good ensemble. In this paper we define a method for introducing diversity among the base learners of one of the previously presented multi-label ensemble classifiers. An empirical comparison on 10 datasets demonstrates that model diversity leads to an improvement in prediction accuracy in 80% of the evaluated cases. Additionally, in most cases the proposed "diverse" ensemble method outperforms other multi-label ensembles as well.

1 Introduction

Multi-label classification problems have attracted many researchers in recent years. A variety of methods especially designed for handling multi-label data have been developed. In addition to the single-model classifiers, a number of ensemble methods have been proposed. Ensemble methods usually improve the prediction performance over a single classifier [1, 2]. It is also well known that the performance of an ensemble is related to both the accuracy and diversity of its base learners. Krogh and Vedelsby [3] have shown that ensemble errors depend not only on the average error of the base models, but also on their diversity.

The questions addressed in this paper are whether if and how a performance of multi-label ensemble methods can be maximized by introducing additional diversity among their models. Specifically, we address the ChiDep Ensemble method proposed in [4] which gathers several different ChiDep classifiers into a composite ensemble model. We have defined a strategy for selecting the more diverse models for participation in the ensemble rather than simply selecting the *m* models with the highest "dependency" score, as was done in the original version of the ensemble in [4]. The defined procedure allows a trade-off between the "diversity" and "dependency" scores among the models, thus providing an opportunity to maximize "diversity" with minimal or no decrease of the model accuracy.

The rest of the paper is structured as follows. In the next section, the original ChiDep ensemble algorithm is briefly described and then the new "diverse" version is proposed. Section 3 presents the setup of the empirical experiment and its results. Finally, Section 4 concludes the current work and outlines some further research directions.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

2 Introducing diversity into ChiDep Ensemble

The basic procedure of the single-model ChiDep classifier is as follows; decomposition of the original set of labels into several non-overlapping subsets of dependent labels, building an Label Power-set (LP) [5] classifier for each subset, and combining them similarly to the Binary Relevance (BR) [5] method. The original ChiDep Ensemble method gathers several different ChiDep classifiers with the highest "dependency" scores into a composite ensemble model. The level of dependency between each two labels in the dataset is identified by applying the chi-square test for independence to the number of train instances for each combination of these two labels. Then, a large number (i.e., 50000) of possible label set partitions is randomly generated and a score for each partition is computed according to the dependence (i.e. χ^2) score of all label pairs within the partition. Subsequently, the top *m* of the high-scored label set partitions are selected as members of the ensemble. A detailed description of the ChiDep classifier and its ensemble can be found in [4].

It is known that the accuracy of an ensemble classifier might be improved by selecting more diverse base learners. In order to achieve additional improvement of ChiDep ensemble accuracy, we have defined a strategy for selecting the most different from the highly scored models for participation in the ensemble. For this purpose we utilize the distance function defined in [6]. For each pair of labels l_i, l_j, the distance function adds "1" to the total distance if both labels belong to the same subset in one partitioning structure and to different subsets in the other partitioning structure. This assures that as more different are the partitioning structures as higher is their distance value. The following procedure was used for selecting the most different (i.e., diverse) partitions from among the highly scored ones.

1. Compute the distance matrix between all pairs of *N* partitions with highest "dependency" scores. Later we will refer to these *N* high-scored partitions as the set of "*candidate*" models.
2. Select the label set partition with the highest dependence score and add it to the set of "*selected*" models for participation in the ensemble.
3. Find the minimal distance $d_{i,min}$ from each one of the "*candidate*" models to all "*selected*" models.
4. Sort all "*candidate*" models in descending order of their d_{min} value (in step 3) and select *k* percent of the partitions with the highest d_{min}. This step is intended to choose *k* percents of the "*candidate*" models that are most different from the "*selected*" models. We refer to this set of models as "*best candidates*".
5. From the "*best candidates*" set, select the partition with the highest dependence score and add it to the set of "*selected*" models.
6. Repeat steps 3-5 until the number of the models in the set of "*selected*" models reaches *m*.
7. Return the "*selected*" set of models for participating in the ensemble.

This procedure allows to trade-off between the "diversity" and "dependency" scores among the selected models. The parameters *N* and *k* influence the model selection process and the level of "diversity"-"dependency" among the ensemble models. *N* defines the number of high-scored partitions which would be considered as "candidates" for ensemble. The higher this number, then the more different they are to

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

each other, however, with lower "dependency" scores, partitions would be selected. For example, for a dataset with 6 labels, there are 172 possible distinct label-set partitions whose dependence score may vary from high positive to high negative numbers. Thus, when defining *N=100*, more than half of all possible partitions will be considered and a portion of them will likely have negative "dependency" scores. However, for a dataset with 14 labels, there are above 6300 possible distinct label-set partitions and it is probable that all of the 100 high scored ones will have high positive "dependency" scores. Thus, for datasets with a small number of labels or low dependency levels among the labels, relatively small values (between 20 and 100) for *N* should be considered. However, for datasets with large numbers of labels or higher dependency levels among the labels, higher values of *N* (100 and above) are likely to perform better. Parameter *k* allows to dynamically define a threshold value for models which are "different enough" from the already selected ones. For example, given that *N=100*, setting *k* to *0.2* signifies that all 20 (20 percent of 100) of the most different models from all the currently selected models will be considered as sufficiently different and, in the end, the one with the highest dependency score will be added to the ensemble. Larger values of *k* are expected to reduce the level of diversity among the selected models, as partitions with lower distance to the "*selected*" models will be considered as sufficiently different and could be also selected to ensemble. Clearly the "best" values for these parameters are dependent on dataset properties. Thus, in order to achieve the best performance, we recommend calibrating these parameters for each dataset specifically.

In this research we carried out a variety of global calibration experiments in order to define the appropriate default values which would allow parameters to perform sufficiently for most datasets. The selected values are presented in the following section.

3 Empirical Evaluation

3.1 Evaluation setup

We empirically evaluated the proposed approach by measuring its performance on ten benchmark multi-label datasets* from different domains and varying sizes. The tests were performed using original train and test dataset splits. We compared the results achieved by the proposed "diverse" ChiDep ensemble to those of the original ensemble version [4] and other state-of-the-art ensemble methods, namely, RAkEL [7], Ensemble of Classifier Chains (ECC) [8] and Ensemble of Pruned Sets (EPS) [9].

All methods were evaluated using the Mulan† library. The "diversity" extension for the ChiDep ensemble was implemented using this library as well. All the algorithms were supplied with Weka's J48 implementation of a C4.5 tree classifier as a single-label base learner. We averaged the results of each of these algorithms over five distinct runs on each dataset. Consecutive numbers from 1 to 5 were used as an initialization seed for the random number generator, allowing for reproducibility of the experimental results.

* The data sets are available at http://mlkd.csd.auth.gr/multilabel.html
† Software is available at http://mulan.sourceforge.net/

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

We compared the algorithms results in terms of the Classification Accuracy evaluation measure [5]. The statistical significance of differences between algorithm results was determined by the Friedman test [10] and post-hoc Holm's procedure for controlling the family-wise error in multiple hypothesis testing.

The number of models participating in the ensemble classifier is expected to influence the predictive accuracy of the classifier. For the sake of fair comparison, we wanted to evaluate ensemble models of an equivalent complexity. To achieve this, we configured all the ensemble algorithms in the experiment in order to construct the same number of distinct models. For the ChiDep ensemble, we set the number of label set partitions *m* to 10 (as is frequently used for the number of classifiers in an ensemble) and averaged the number of distinct models constructed by the ensemble across all random runs. This number, presented in Table 2, in the *CDEd-Models* column, was taken as the base number of models for all ensemble methods. RAkEL, ECC and EPS were configured to construct the same number of distinct models. For all ensemble methods, the *majority voting threshold* was set to a commonly used intuitive value of *0.5*.

Other parameters of ensemble methods were configured as follows. The RAkEL's *k* parameter was set to 3. ECC does not require additional parameters. For the EPS, *p* and *s* parameters were set to the values chosen by the authors for the datasets presented in the PS paper [9]. For other datasets, we set *p=1* (as the dominant value among chosen values in the PS paper) while *s* was computed by PS's utility (from Meka[‡] library), as recommended by the authors. The "diverse" version of the ChiDep ensemble was supplied with *N=100* and *k=0.2*, accordingly to the results of the calibration experiments described in Section 3.2.

3.2 **Experimental results**

First, we compare the predictive performance of the proposed "diverse" version of the ChiDep Ensemble (CDE-d) method to its original "base" version (CDE-b). According to the "diverse" version of the ChilDep ensemble we try to select the most different (at least in *k* percent) models among the *N*-high scored ones. The *k* and *N* are configured parameters and might be specific for each dataset.

We performed several calibration experiments to examine whether the "diverse" version of the CDE algorithm in combination with default values for the configured parameters can improve the predictive performance of the "base" version. The calibration experiments were run on scene, emotions, yeast and medical train sets using 10-fold cross validation with parameter *k* varying from 0.1 to 0.9 with step 0.1 and parameter *N* varying from 100 to 500 with step 50. In the result analysis, we found that combination of values *k=0.2* and *N=100* performed well and was the only combination appearing among the 25 best results on all evaluated datasets. The test of CDE-d with the selected parameters on all the datasets showed that indeed the "diverse" version of the CDE algorithm, even with default parameters, improves the predictive performance of the ensemble.

Table 1 presents the results of the CDE-d algorithm with the selected default parameters on all datasets and compares them to those of the base CDE version. As

[‡] Software is available at http://meka.sourceforge.net/

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

observed from the table, model diversity leads to an improvement of prediction accuracy for the ensemble classifier in 8 out of 10 evaluated datasets. Note that higher predictive performance can be achieved by specifically calibrating the parameters for each dataset.

dataset	CDE-d	CDE-b
emotions	**53.92**	51.05
scene	**60.05**	59.91
yeast	**49.7**	49.04
genbase	**98.66**	98.56
medical	**71.56**	71.48
enron	**43.26**	43.14
tmc2007-500	83.55	**83.7**
rcv1(subset1)**	**7.2**	**7.2**
mediamill	**43.11**	42.54
bibtex	**30.17**	30.06

Table 1: Classification accuracy of CDE "diverse" and "base" versions.

Next, let us compare the CDE-d to other multi-label ensemble algorithms. The results of the evaluated ensemble classifiers are presented in Table 2. The differences between algorithms were found to be statistically significant by the Friedman test, at a significance level of $p=0.02$. The followed post-hoc Holm's procedure indicates that there is no significant case where RAkEL, ECC or EPS is more accurate than CDE-d. On the other hand, CDE-d is significantly more accurate than EPS. In addition, the CDE-d accuracy values are higher than those of ECC and RAkEL algorithms in most cases; however, these differences are not statistically significant. In general, the CDE-d method obtains the best average rank among all the evaluated methods. We can also observe that CDE-d algorithm achieves the best results on 6 out of 10 datasets.

dataset	CDEd-Models	CDEd-Rank	CDE-d	RAkEL	ECC	EPS
emotions	22	2	53.92	51.28	**54.02**	53.63
scene	21	3	60.05	60.87	**62.24**	58.5
yeast	37	1	**49.7**	48.39	45.63	49.14
genbase	111	1.5	**98.66**	**98.66**	98.63	98.32
medical	239	3	71.56	71.14	73.82	**75.03**
enron	249	1	**43.26**	41.89	42.26	34.65
tmc2007-500	77	1	**83.55**	80.90	72.19	75.69§
rcv1(subset1)**	473	1	**7.2**	7.04	6.662	6.94§
mediamill	459	2	43.11	**44.52**	40.89	OOM††
bibtex	761	1	**30.17**	29.88	29.37	OOM††
Avg. rank	-	-	**1.7**	2.5	2.7	3.0

Table 2: Classification accuracy of CDE-d and other multi-label ensemble algorithms.

§ The result is for the 10 models ensemble, due to Java OutOfMemory exception with the specified number of models.

** The version of dataset with 944 features was used, as in [11].

†† The ensemble caused OutOfMemory Exception even with 10 base-learning models.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

4 Conclusions

In this paper we have presented a novel method for introducing diversity among the base learners of the known multi-label classification ensemble ChiDep. We evaluated the new algorithm on 10 datasets of various complexities.

First, we compared the new "diverse" version of the ChiDep ensemble to the original one. The results of this comparison confirm that the models' diversity improves ensemble's prediction performance over its original version. We then compared the new algorithm to three other known multi-label ensemble classifiers. The results demonstrate that the proposed new algorithm is able to improve prediction accuracy over other well known multi-label classification ensembles as well.

Summarizing the results of this evaluation experiment, we conclude that introducing additional diversity among the models of multi-label ensembles allows for maximization of their performance. In the future we are planning to verify if using the diversity criteria alone (instead of the trade-off between "dependency" and "diversity") would improve, even more, the ChiDep ensemble accuracy. Moreover, exploration and development of standard methods for introducing diversity into other multi-label ensemble methods is among additional issues to be further studied.

References

[1] T. G. Dietterich, Ensemble Methods in Machine Learning. Proc. First International Workshop on Multiple Classifier Systems, J. Kittler and F. Roli, Eds. New York: Springer Verlag, pp. 1-15, 2000.

[2] L. Rokach and O. Maimon, *Feature set decomposition for decision trees*. Journal of Intelligent Data Analysis, 9(2):131–158.

[3] A. Krogh and J. Vedelsby, Neural network ensembles, cross validation, and active learning. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, Advances in Neural Information Processing Systems, MIT Press, pp. 231–238, 1995.

[4] L. Tenenboim-Chekina, L. Rokach, B. Shapira, Identification of Label Dependencies for Multi-label Classification, Proc. of ICML Workshop on Learning from Multi-Label Data, pp. 53-60, 2010.

[5] G. Tsoumakas, I. Katakis, and I. Vlahavas, Mining Multi-label Data. In O. Maimon and L. Rokach, editors, Data Mining and Knowledge Discovery Handbook, New York: Springer, pp. 667-686, 2010.

[6] L. Rokach, *Genetic algorithm-based feature set partitioning for classification problems*. Pattern Recognition, 41(5):1693-1717, 2008. doi=http://dx.doi.org/10.1016/j.patcog.2007.10.013

[7] G. Tsoumakas and I. Vlahavas, Random k-Labelsets: An Ensemble Method for Multilabel Classification. Proc. of 18th European Conference on Machine Learning, pp. 406-417, 2007.

[8] J. Read, B. Pfahringer, G. Holmes, and E. Frank, Classier Chains for Multi-label Classification. Proc. of ECML-KDD, 2, pp. 254-269, 2009.

[9] J. Read, B. Pfahringer, G. Holmes, Multi-label classification using ensembles of pruned sets. Proc. of Eighth IEEE International Conference on Data Mining, 0, 995-1000, 2008.

[10] J. Demsar, *Statistical comparisons of classifiers over multiple data sets*. Journal of Machine Learning Research, 7:1-30, 2006.

[11] M. Zhang, and K. Zhang, Multi-label learning by exploiting label dependency. In Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 999-1008, Washington, DC, USA. doi= http://doi.acm.org/10.1145/1835804.1835930