Evaluation of Fusion Approaches in Large-scale Bio-annotation Setting*

Vedrana Vidulin[1§], Maria Brbić[1] Fran Supek[1,2], and Tomislav Šmuc[1]

[1] Ruđer Boškovic Institute, Bijenička cesta 54, 10000 Zagreb, Croatia
[§]vedrana.vidulin@irb.hr,
[2] EMBL/CRG Systems Biology Unit, Centre for Genomic Regulation, Dr. Aiguader 88, 08003 Barcelona, Spain

Abstract. In this work we compare different information fusion approaches in the context of large-scale multi-label classification problems, typical today in bio-domains: early fusion, late fusion and hybrid fusion approach. The experiments are performed on two novel large-scale classification datasets for gene function prediction and prokaryotic phenotype prediction. Both datasets are based on descriptors coming from a number of different representations of biological entities. The results reveal that the fusion approaches exploiting complementarity are best suited for difficult annoation problems featured in complex datasets from bio-domains for which individual classifiers perform well only locally.

Keywords: information fusion, ensemble classifier design, diversity, annotation, genomics

1 Introduction

Combining classifiers for information fusion [1], [2], [3] is an important topic in the era of information overload present in the majority of technological and scientific domains, from multimedia systems to life-sciences. In particular, in molecular and systems biology multitude of omics approaches are used in order to explain complex roles and associations between cell constituents, and machine learning methods are used as tools for knowledge discovery, either for annotating entities with typically non-exclusive roles (e.g. protein function prediction), or in searching for the important patterns, interactions and representations of entities for the particular problem at hand. The importance of machine learning is best seen through a number of predictive challenges, some of which are repetitive regular events such as CASP and CAFA. Here are some features accompanying discovery problems in genomics:

1. Predictive performance is typically optimized to maximize predictions at certain precision threshold, and the final model can refrain from making a

* The first two authors should be regarded as joint first authors.

decision when certain precision is not guaranteed. This predictive setting is actually close to that of information retrieval in which metrics uch as F-measure, Area Under Precision Recall Curve (AUPRC) or recall@precision are used.
2. Predictive problems often involve multi-label classification or structured output prediction, where the output structure can be in the form of taxonomy or ontology showing the need for tools that exploit relations between labels.
3. Complex problems in this domain often include a number of different representations drawn from at least partially independent views. Overall dimensionality of these representations, in terms of the number of features, their relevance for the target and their mutual independence are the major challenges with respect to the annotation problems. Therefore, it is important to have efficient and reliable methods that can exploit individual contexts and their interactions.

The last feature can be handled using classifier fusion approaches that exploit complementarity on a feature or classifier level. In this work we evaluate different fusion approaches in the context of two large scale predictive problems: gene function prediction (GFP) and prokaryotic phenotype prediction (PP), addressing all of the specific requirements mentioned above. Evaluation methodology provide us a detailed picture of driving mechanisms behind the predictive performance of different approaches. The main contribution of this work are insights related to the importance of fusion mechanisms and their capability to exploit advantages of different representations used to describe entities of underlying problems.

2 Related Work

Multi-classifier systems, classifier ensembles and meta-learning approaches have been important topics in the machine learning field for several decades. Information fusion is closely related to these topics. The evolution and adaptation of the well known techniques and approaches in this field is of high importance in the context of complex, distributed and streaming data environments. Concepts of fusion and multi-classifier combination techniques are covered extensively in the reviews [1], [3], [4], and have been important topics at the relevant conferences in the field of data mining and machine learning.

Different aspects are important when optimizing combination of modalities, but one most considered is the level at which fusion of representations is performed: in early fusion predictive model is built using all feature sets or representations together to make a single decision model; contrary to that, late fusion approaches deal with combining models learnt separately on different features sets or representations. Each of these approaches has their own advantages. Early fusion approach can make use of interactions between basic features from different representations, improving both understanding and predictive performance, while late fusion approaches can combine outputs of different classifiers exploiting the same representation on a decision level. Moreover, combining individual

decisions from possibly complex individual representations offers scalability and allows us to use the most suitable methods for analyzing each single representation, thus providing more flexibility than the early fusion.

An obvious extension that combines the strengths of both approaches is the hybrid fusion, which in the simplest variant, combines early fusion model with the individual representation models using late fusion approaches.

In the majority of analysis reported in literature empirical results demonstrate better performance of late fusion methods in comparison to early fusion. Late fusion methods exploit additional level of complexity through another layer of decision making, i.e. combination of multiple baseline (feature subset) models. Although the power of ensembles approaches has been both theoretically and empirically proven to be related to at least partial independence of models, the strong and definite connection between different ensemble diversity and accuracy of the ensemble seems to be lacking [5].

Recently, in paper by [6] the authors illustrate the connection between improvement of ensemble classifiers based on the relative independence of their false-positive prediction patterns. In their study they used recall at predefined precision level as a measure to compare classifiers' performances. This measure is very well suited for information retrieval, but it is also aligned with the knowledge discovery tasks. We show in this work that the power of late fusion approaches is most effective in the setting where strong predictive performance of individual classifiers (or descriptor sets) is of very local character. Furthermore, we show that in this setting that there is a strong correlation between recall at predefined precision threshold and the diversity of classifiers calculated using their individual performances.

3 Fusion approaches

In the early fusion setting feature sets from the individual representations are all joined into one dataset from which a single classifier is constructed. This type of fusion should generally better exploit interactions between features from different representations.

On the other hand, late fusion is performed by constructing a separate classifier for each of the individual representations and then fusing their predictions. Our pipeline implements five different late fusion approaches: one vote, two votes, three votes, consensus [7] and weighted voting [8].

Let $C = (c_i)_{i=1}^{N}$ be a sequence of confidence scores of N individual classifiers and let $S = (s_i)_{i=1}^{N} = sort(C)$ denote a sequence arranged in ascending order i.e. $s_i \leq s_{i+1}$. The one/two/three votes approaches calculate the fused confidence scored of the class y_j:

$$c_{1vote}(y_j) = s_N(y_j); \quad c_{2votes}(y_j) = s_{N-1}(y_j); \quad c_{3votes}(y_j) = s_{N-2}(y_j) \tag{1}$$

Consensus score of a label y_j is calculated using the following formula:

$$c_{\text{cons}}(y_j) = 1 - \prod_{i=1}^{N}(1 - c_i(y_j)) \tag{2}$$

The weighted voting score of a label y_j is calculated as follows:

$$c_{\mathrm{wv}}(y_j) = \sum_{i=1}^{N} w_i c_i(y_j), \tag{3}$$

where w_i denotes weight assigned to the classifer i and:

$$\sum_{i=1}^{N} w_i = 1 \tag{4}$$

We calculated weight as Area Under Precision-Recall Curve (AUPRC) estimated in the cross-validation setting and normalized to sum to 1.

Finally, hybrid fusion performs late fusion on the predictions from the individual classifiers and the early fusion classifier.

4 Experimental Setup

In this section we start by introducing problems and respective representations used in this work. Further, we explain the general scheme of experiments designed to give us an answer to our main question: which fusion approaches work best for the types of problems we solve in this work, while also providing intuition when and why. Next, we present the experimental methodology and discuss the evaluation measures used in the experiments.

4.1 Datasets

Gene function prediction. Five datasets represent GFP methods based on semantically distinct feature sets (Table 1, details in [9]). All datasets have common set of 15,308 instances representing eggNOG clusters of genes [10] and are labeled with 935 gene functions from Gene Ontology (GO) [11].

Phenotype prediction. Phenotype prediction datasets are constructed using six different representations (Table 2, details in [12]). Each representation represents one dataset with a set of 703 instances and 72 labels. Each instance corresponds to one prokaryotic organism labelled with a set of phenotypic traits.

4.2 Methodology

The input to our computational pipeline is the group of datasets that describe the same concept but with distinct feature sets. First, in order to reduce the dimensionality of the feature sets we applied principal component analysis (PCA) on each of the datasets separately and retained principal components (PC) that explain 90% of the variance. We divided our data into training (consisting of 2/3 learning instances) and test sets (1/3 instances) using stratified sampling.

Table 1. Gene function prediction datasets.

Dataset	# features	# PC	Features description
Phyletic profiles (PP)	2071	352	Features are genomes and feature values represent presence/absence of cluster member genes in genomes.
Empirical kernel map (EKM)	8447	1552	Features are gene clusters and feature values represent minimal distance between gene sequence pairs, where one sequence is from instance and another from feature cluster. Distance is measured as e-value.
Conserved gene neighborhood (CGN)	5891	1411	Features are gene clusters and feature values represent average log-distance (in nucleotides) between genes from instance and feature cluster pairs averaged over genomes.
Translation efficiency profiles (TEP)	2071	1449	Features are genomes and values represent maximal predicted level of cluster member genes expression.
Biophysical and protein sequence properties (BPS)	1170	296	Features represent various gene sequence statistics as described in [13]. Statistics are computed on a gene level and averaged between cluster member genes.

Table 2. Phenotype prediction datasets.

Dataset	# features	# PC	Features description
Text-mining	95663	438	Bag-of-words representations of documents describing bacterial/archael species collected from the scientific literature and the broader World Wide Web
Amino acid content	420	3	Amino acid and di-amino acid frequencies of a proteome
Pairwise co-occurrences	1235	33	Pairwise co-occurrences of species in metagenomes [14]
Phyletic profiles	80576	393	Presence/absence of the clusters of orthologous groups (COGs) of proteins
Conserved gene neighborhood	44850	366	Log pairwise chromosomal distance in nucleotides between pairs of 300 frequently occurring COG gene families
Translation efficiency profiles	990	263	Codon usage biases of COG gene families across 606 genomes, measured using the MILC method [15]

In the early fusion setting, all feature sets were combined and given as an input to a single early fusion model (EF). In the late fusion setting, individual models are constructed for each of the feature sets (FS) separately. In order to access performance of indvidual classifiers necessary for the weighted voting approach, we used a cross-validation setting. The hybrid fusion model was built in the same way as the late fusion one, but also using the early fusion model as a classifier. Finally, all models are deployed on the test set instances. Figure 4.2 gives a general outline of computational experiments performed on each of the problems discussed in this work.

Governed by the principle of the best trade-off of accuracy, efficiency and robustness, we used random forests [16] as a classification algorithm. Random forests have been used on biological data and the evaluation shows that they are able to produce state-of-the-art annotation results [9] [23] [24]. Since label spaces of gene function prediction (GFP) and phenotype prediction (PP) problems have different properties, we used different versions of the algorithm. In order to exploit hierarchical structure of the labels in the GFP problem, we used random forests version of CLUS-HMC [17], [18]. CLUS-HMC is the algorithm for hierarchical multilabel classification based on the framework of Predictive Clustering Trees [19]. CLUS-HMC was run with the default parameters, except

for these settings: decision tree pre-pruning was used to prevent the algorithm to form a leaf node when the number of instances in the node is <5; forest size was set to 200 trees; size of a feature subset for random forests was set to square root of the total number of features. For the PP problem where the multi-label target side is flat, we used FastRF, a fast and efficient implementation of the random forest algorithm in WEKA [20]. The number of trees was set to 500. For this setting we applied binary relevance method that corresponds to learning one classifier for each label separately. This leads to a notable difference from CLUS-HMC which is able to produce one global model over the whole hierarchy of labels. Finally, it is important to note that diversity of individual models in the fusion ensemble is the consequence of the use of different feature sets given in Tables 1 and 2 in building individual classifiers of the fusion ensemble.

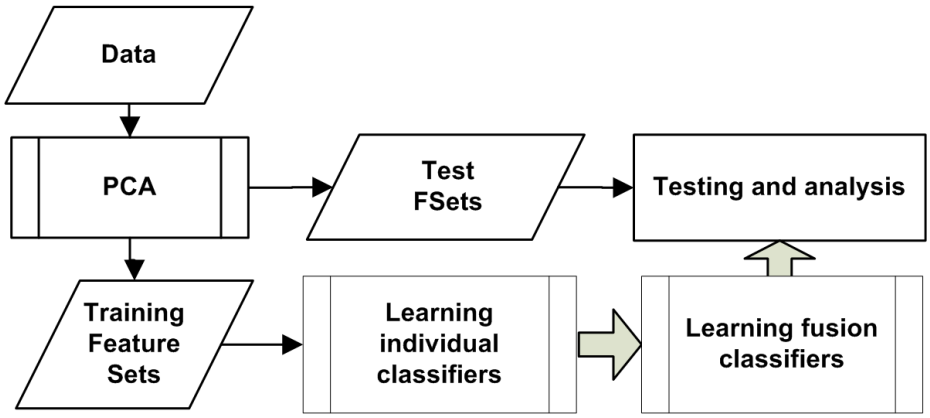

Fig. 1. General schema of computational experiments performed in this work.

4.3 Evaluation

Predictive performance of different fusion approaches was measured on a separate test set, composed of one third of the instances. We used two performance measures that rely on a precision-recall curve: (i) Area Under Precision Recall Curve (AUPRC), and (ii) recall at some predefined precision level (R@P) which represents the part of the precision-recall curve at some (high) precision level. The latter measure was used to emphasize the importance of having predictions with high level of precision which is especially important in the context of annotation for omics data.

A single PR curve was computed by averaging label-specific curves, which corresponds to the averaging procedure known as macro-averaging in a multi-label machine learning setting. It is common to report micro beside macro-averaged measures, but micro-averaging is not appropriate in the setup with

highly unbalanced classes where interesting classes are typically those with the least positive examples. For example, a specific label predicted from the bottom of the GO ontology is more interesting for a domain expert than the label predicted higher in the ontology. In such settings, micro-averaging would equally weight examples and thus, averaged performance scores would mainly represent the performance on less interesting general labels. In contrast, macro-averaging equally weights labels, enabling interesting specific classes to influence the average performance score.

Statistical comparison of fusion approaches was performed by using the corrected Friedman test and the post-hoc Nemenyi test [25]. The Friedman test is a non-parametric test for multiple hypothesis testing. For each label, this test ranks the fusion approaches according to AUPRCs measured for this specific label. The best approach is ranked as the first and in the case of ties an average rank is assigned. The test compares approaches by comparing ranks averaged over all labels and calculates Friedman statistic distributed according to the χ^2_F with $k-1$ degrees of freedom, k being the number of fusion approaches. In cases where at least one approach performed significantly different than the rest we performed Nemenyi test that shows where that difference lies. We present the results of Nemenyi test using average ranks diagrams [25].

Diversity in classifier ensembles is measured as a disagreement between classifiers in terms of correct/incorrect predictions. Different pairwise diversity measures have been proposed in literature [5], while the averaged statistic is typically calculated by averaging over all pairs of classifiers. As the most appropriate measure to characterize diversity in our setting, we chose the disagreement measure [21] defined as the ratio between the number of predictions on which classifiers disagree and the total number of predictions:

$$Disagreement(i,j) = \frac{N^{10} + N^{01}}{N^{11} + N^{10} + N^{01} + N^{00}}, \tag{5}$$

where N^{10} denotes the number of predictions on which classifer i is correct and classifier j is incorrect; the same applies vice versa. Since we deal with highly unbalanced classes, the diversity measure was calculated only on positive examples.

In order to investigate the complementarity in the context of fusion performance in more detail, we assessed the performance of fusion approaches in the respect to the generality of labels. The generality was measured using information content (IC) [22] computed from label frequency. Higher IC is related to the more specific labels which are usually more difficult to predict, but more valuable to the domain experts.

4.4 Experiments

In order to investigate which of the fusion approaches performs best and how is the performance related to the diversity of individual classifiers, we structured our experiments in the following manner: (i) we computed macro-averaged performance measures for each of the individual classifiers and fusion approaches

across three levels of difficulty; (ii) we looked at the relationship between difficulty and diversity of individual classifiers for EF, LF and HF approaches; and (iii) to examine complementarity of EF, LF and HF we measured improvement over a baseline. Since in our experiments LF-three votes approach is a proxy of majority voting and a solution that fosters consensus of classifiers rather than complementarity, we used it as a baseline.

5 Results and discussion

Performances of individual classifiers and different fusion approaches for GFP and PP problem are shown in Table 3 and Table 4, respectively. Performances are measured using average AUPRC and recall at 70% precision threshold. Results are structured to show performance for different label generality levels defined by problem the specific IC intervals (general, medium and specific; equal number of labels per bin). One obvious difference between the GFP and PP problems is that the average performance of individual classifiers for the GFP problem is much lower than for the PP problem. This is not unexpected since the GFP annotation is more difficult problem with more than an order of magnitude larger label space. Average results in Tables 3 and Table 4 are accompanied with statistical comparison of fusion approaches through average ranks diagrams across label generality levels shown in Figure 2.

More detailed analysis reveals different trends between fusion approaches across label generality levels. For the GFP problem, weighted voting approach and consensus scheme are clear winners over all generality levels, while late fusion and hybrid fusion approaches are consistently higher than the early fusion approach. The situation is different for the PP problem: early fusion gives slightly better results than late fusion and hybrid fusion approaches. For the general and medium-specific labels of the PP problem, three votes fusion seems to be a better choice than the weighted voting or consensus scheme, albeit not significantly. Also, differences in ranking between one and three vote schemes for the two problems suggest that the improvements through fusion is obtained in a different manner: by exploiting complementarity in GFP, and consensus in PP problem. In relative terms, improvements obtained by fusion approaches are much more significant for the GFP problem, as are also the differences between fusion schemes, as depicted in Figure 2.

5.1 Classifier diversity and performance of fusion approaches

The relationship between diversity and performance of different fusion approaches is shown on Figure 3. For both problems, higher diversity seems to be clearly related with higher performance, similarly for AUPRC and R@P measure. It seems that diversity spreads to larger values for GFP than for PP problem. Before making further inferences about the nature of the relationship between performance of classifiers and diversity we need to look into basic characteristics

Table 3. Gene function prediction

		GENERAL		MEDIUM-SPECIF		SPECIFIC	
		AUPRC	RC@0.7PR	AUPRC	RC@0.7PR	AUPRC	RC@0.7PR
Late fusion	Early fusion	0.543	0.417	0.196	0.192	0.081	0.155
	One vote	0.573	0.440	0.238	0.201	0.113	0.149
	Two votes	0.453	0.335	0.176	0.170	0.103	0.147
	Three votes	0.241	0.127	0.118	0.100	0.044	0.096
	Consensus	0.571	0.447	0.252	0.227	0.131	0.169
	Weighted voting	0.590	0.455	0.257	0.228	0.122	0.165
Hybrid fus	One vote	0.575	0.439	0.245	0.201	0.117	0.151
	Two votes	0.561	0.434	0.251	0.216	0.131	0.175
	Three votes	0.485	0.373	0.210	0.200	0.107	0.165
	Consensus	0.582	0.449	0.275	0.233	0.153	0.187
	Weighted voting	0.580	0.453	0.276	0.235	0.134	0.186
Individual	PP	0.136	0.048	0.057	0.065	0.007	0.065
	EKM	0.529	0.392	0.144	0.163	0.030	0.102
	CGN	0.123	0.046	0.045	0.062	0.008	0.075
	TEP	0.083	0.026	0.018	0.044	0.004	0.026
	BPS	0.411	0.264	0.052	0.089	0.023	0.104

Table 4. Phenotype prediction

		GENERAL		MEDIUM-SPECIF		SPECIFIC	
		AUPRC	RC@0.7PR	AUPRC	RC@0.7PR	AUPRC	RC@0.7PR
Late fusion	Early fusion	0.836	0.781	0.610	0.465	0.439	0.309
	One vote	0.732	0.622	0.452	0.227	0.373	0.227
	Two votes	0.796	0.710	0.562	0.413	0.434	0.312
	Three votes	0.820	0.774	0.571	0.430	0.425	0.333
	Consensus	0.799	0.726	0.551	0.383	0.406	0.265
	Weighted voting	0.820	0.762	0.584	0.436	0.430	0.296
Hybrid fus	One vote	0.732	0.623	0.452	0.227	0.373	0.227
	Two votes	0.802	0.711	0.581	0.439	0.440	0.312
	Three votes	0.831	0.787	0.595	0.449	0.439	0.319
	Consensus	0.808	0.736	0.566	0.399	0.416	0.279
	Weighted voting	0.827	0.772	0.593	0.445	0.446	0.317
Individual	TM	0.791	0.715	0.575	0.428	0.388	0.262
	AAC	0.679	0.540	0.339	0.153	0.270	0.126
	PC	0.611	0.408	0.253	0.119	0.132	0.058
	CGN	0.759	0.692	0.499	0.361	0.429	0.336
	PP	0.781	0.696	0.504	0.347	0.414	0.342
	TEP	0.752	0.684	0.418	0.240	0.259	0.177

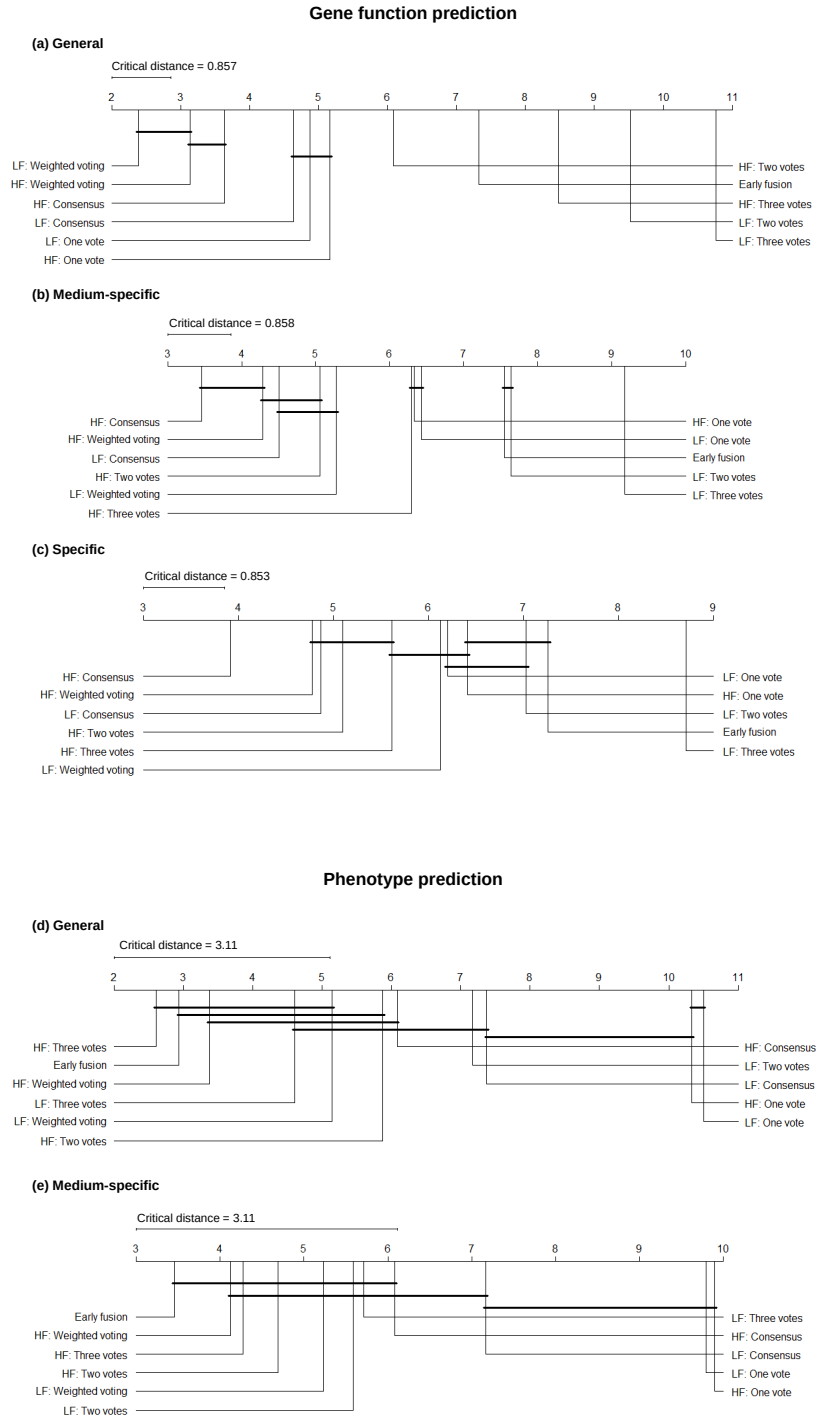

Fig. 2. Average ranks diagrams comparing the performance of fusion approaches over labels belonging to different levels of generality. LF stands for late fusion and HF for hybrid fusion. The numbers on the axis represent ranks and the best ranking approaches are at the leftmost side of the diagram. The approaches that do not differ by less than the critical distance for a significance level of 0.05 are connected with a line.

of the used diversity measure. Assuming independence between individual classifier outputs, higher values of diversity can be expected for ensembles of medium performing classifiers or ensembles with very diverse range of performances. If we assume that we have homogeneous ensemble of classifiers with high recall values, we can expect low diversity measures. The same is true if all classifiers have low recall values since in that case low diversity is driven by high values of N^{00} in Eq.(5). Main correlation trend between diversity and performance for the GFP problem seems to be driven by ensembles of low performing - low diversity classifiers on one hand and mixtures of ensembles with low and high performing classifiers (range of diversity 0.3-0.6).

The same mixture of cases seems to be driving the correlation behind performance measures and diversity in case of the PP data, with one notable difference, the cases with low diversity and very high score. These cases together with the rest of general labels exhibit no correlation between diversity and performance scores.

5.2 Relationship between diversity and generality of labels

In principle, when dividing the label space, we assumed that more general labels will be easier to learn and this holds on average for both problems. However, for GFP problem in particular, diversity seems to play important role for general labels as well. At the first glance there is a counter-intuitive result for the GFP problem: the correlation of diversity and performance is somewhat larger for general categories than for the specific ones. However, there is a considerable number of specific categories that are practically not learnable (R@P=0) at all, which reduces correlation over learnable specific categories. Low learnability seems also to be the explanation for low scores of significant part of general and medium specific GFP labels, which shows that available GFP feature sets are still not fit enough for proper learning of significant part of Gene Ontology graph.

To the contrast, for the PP problem there seems to be no significant correlation between diversity and performance for general category of labels. However, if we compare IC intervals of GFP and PP general category of labels, it can be seen that PP general labels are much more frequent, and therefore also more easily learnable. Overall fitness of feature sets for general PP labels is confirmed through low diversity and very high performance scores for general PP labels. Figure 4 shows increase of the performance scores of fusion approaches over the baseline three votes approach, regarded in our experiment as a proxy for majority voting. Majority voting is the approach that exploits consensus rather than the complementarity of classifiers. On the other hand, weighted voting approach exploits both complementarity and consensus of classifiers. Notable difference between two problems is that improvement for GFP problem is on average significant over all label generality levels, while for the PP problem difference between fusion approaches is practically negligible. Also, for the medium and specific labels of the GFP problem, there is a consistent difference between early, late and hybrid fusion with early fusion as the weakest and hybrid fusion as the

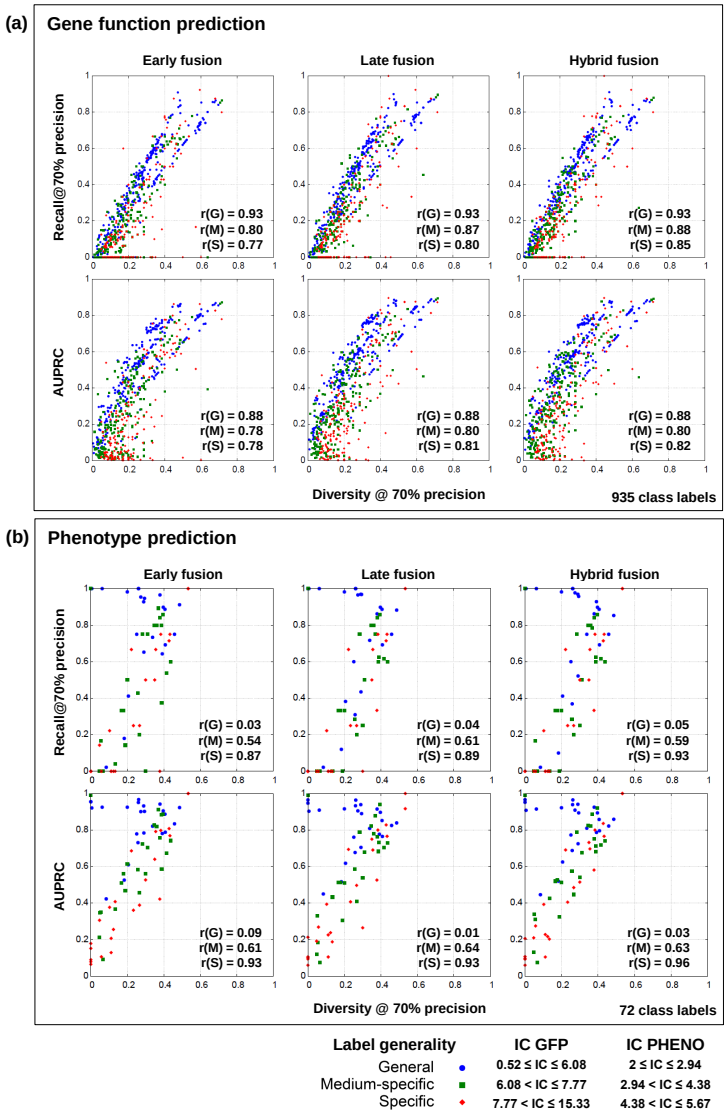

Fig. 3. Diversity of individual classifiers is correlated with the predictive performance of fusion approaches. Late and hybrid fusion are both based on weighted voting approach. r stands for Pearson correlation coefficient, G for general labels, M for medium specific and S for specific labels. IC stands for information content.

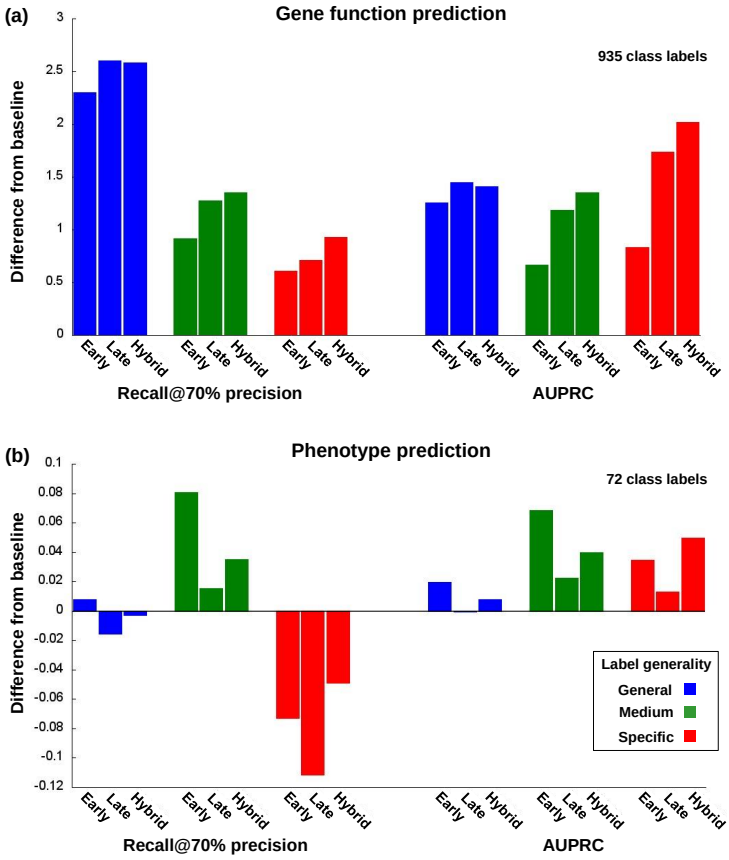

Fig. 4. Differences between fusion approaches and three votes baseline. Late and hybrid fusion results are those for Weighted voting approach.

strongest performer. The logical conclusion is that fusion approaches exploiting complementarity are the best choice for difficult annotation problems, for which individual classifiers perform well only locally. Such cases are characterized by strong correlation between diversity and performance of fused ensembles.

6 Conclusions

This work revisits the problem of classifier fusion with intention to provide new insight into relationship of classifier diversity and performance of classifier fusion approaches. We have used simple fusion approaches that foster complementarity between classifiers, and two performance measures well suited for discovery setting of biological annotation. Although we have performed the analysis on just two datasets, we believe that their complexity and characteristics provide

enough grounds to conclude that diversity of ensembles is indeed strongly related to improved performance of fusion approaches exploiting complementarity. This is the case when individual classifiers exhibit strong performances only locally (i.e. only for a subset of labels). This conclusion slightly departs from the previous findings [5]. In our future efforts we plan to investigate relationship between diversity and performance of fused ensembles in more detail, over more datasets, using larger number of classifiers, and more complex fusion approaches.

Acknowledgements. This work was supported by the European Commission through the projects MAESTRA [ICT-2013-612944], InnoMol [FP7-REGPOT-2012-2013-1-316289] and project Machine Learning Algorithms for Insightful Analysis of Complex Data Structures of the Croatian Science Foundation [Pr. no. 9623].

References

1. Kuncheva, L.: Combining Pattern Classifiers: Methods and Algorithms.Wiley Interscience, 2004
2. Tulyakov, S., Jaeger, S.,Govindaraju, V., Doermann, D.: Review of Classifier Combination Methods, Studies in Computational Intelligence, Vol 90, pp. 361-386, 2008
3. Wozniak, M.,Graña, M.,Corchado, E.: A survey of multiple classifier systems as hybrid systems. Information Fusion 16, pp. 3-17, 2014
4. Tulyakov, S., Jaeger, S.,Govindaraju, V., Doermann, D. : Review of Classifier Combination Methods. Chapter Machine Learning in Document Analysis and Recognition Vol 90, Studies in Computational Intelligence, pp. 361-386, 2008
5. Kuncheva, L.,Whitaker, C.: Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy Machine Learning, Vol 51, pp. 181-207, 2003
6. Madani, O., Georg, M., Ross, D. A.: On using nearly-independent feature families for high precision and confidence. Machine Learning , V 92 (2), pp. 457-477, 2013
7. Piovesan, D., Giollo, M., Leonardi, E., Ferrari, C., Tosatto, S. C.: INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity. Nucleic acids research, 43(W1), 2015
8. Zhou, Z. H.: Ensemble methods: foundations and algorithms. CRC press, 2012
9. Vidulin,V., Šmuc, T., Supek, F.: Extensive complementarity between gene function prediction methods. Bioinformatics, 2016, doi:10.1093/bioinformatics/btw532
10. Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., ... , Bork, P.: eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic acids research, 2013
11. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ..., Sherlock, G.: Gene Ontology: tool for the unification of biology. Nature genetics, 25(1), pp. 25-29, 2000
12. Brbic, M., Piskorec, M., Vidulin, V., Krisko, A., Supek, F.: The landscape of microbial phenotypic traits and their genetic basis, Submitted, 2016
13. Ofer, D., Linial, M.: ProFET: Feature engineering captures high-level protein functions. Bioinformatics, 2015

14. Chaffron, S., Rehrauer, H., Pernthaler, J. Mering, C. von. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Research 20, pp. 947âĂŞ959, 2010
15. Supek, F. Vlahovicek, K: Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics 6 (182), 2005
16. Breiman, L.: Random forests.Machine Learning, 45 (1), pp. 5-32, 2001
17. Kocev, D., Vens, C., Struyf, J., Dzeroski, S.: Tree ensembles for predicting structured outputs. Pattern Recognition, 46(3), pp. 817-833, 2013
18. Vens, C., Struyf, J. Dzeroski, S., and Blockeel, Hendrik: Decision trees for hierarchical multi-label classification. Machine Learning, 73(2), pp. 185-214, 2008
19. Blockeel, H.: Top-down induction of first order logical decision trees. Ph.D. thesis, Katholieke Universiteit Leuven, Leuven, Belgium, 1998
20. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update; SIGKDD Explorations, 11 (1), 2009
21. Skalak, D.: The sources of increased accuracy for two proposed boosting algorithms. In Proc. American Association for Artificial Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop, 1996
22. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, San Francisco, CA, pp. 448-453, 1995
23. Škunca, N., Bošnjak, M., Kriško, A., Panov, P., Džeroski, S., Šmuc, T., Supek, F. (2013). Phyletic profiling with cliques of orthologs is enhanced by signatures of paralogy relationships. PLoS Comput Biol, 9(1), e1002852
24. Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Kocev, D., Džeroski, S. (2010) Predicting gene function using hierarchical multi-label decision tree ensembles, BMC Bioinformatics, 11(2), 1-14
25. Demšar, J. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7:1âĂŞ30, 2006