Functional Inference in FunCat through the Combination of Hierarchical Ensembles with Data Fusion Methods

Nicolò Cesa Bianchi CESA-BIANCHI@DSI.UNIMI.IT
Matteo Re RE@DSI.UNIMI.IT
Giorgio Valentini VALENTINI@DSI.UNIMI.IT
DSI, Università degli Studi di Milano, via Comelico 39/41, I-20135 Milano, Italia

Abstract

The multi-label hierarchical prediction of gene functions at genome and ontology-wide level is a central problem in bioinformatics, and raises challenging questions from a machine learning standpoint. In this context, multi-label hierarchical ensemble methods that take into account the hierarchical relationships between functional classes have been recently proposed. Various studies also showed that the integration of multiple sources of data is one of the key issues to significantly improve gene function prediction. We propose an integrated approach that combines local data fusion strategies with global hierarchical multi-label methods. The label unbalance typically occurring in gene functional classes is taken into account through the use of cost-sensitive techniques. Ontology-wide results with the yeast model organism, using the FunCat taxonomy, show the effectiveness of the proposed methodological approach.

1. Introduction

Gene function prediction is a complex multi-label classification problem characterized by functional classes structured according to a predefined hierarchy —for example, a directed acyclic graph (The Gene Ontology Consortium, 2000) or a forest of trees (Ruepp et al., 2004). This hierarchy typically contains hundreds or thousands of nodes, and the complexity of the classification problem is further increased by a consistent unbalance between positive and negative examples, and by the need of integrating different types of data to increase the reliability of the functional classification.

In the literature, many approaches have been proposed to deal with the integration of multiple sources of data. Functional linkage networks (Karaoz, 2004), kernel fusion (Lanckriet et al., 2004), vector space integration (Pavlidis et al., 2002) and ensemble systems (Re & Valentini, 2010) have been proven their effectiveness for gene function prediction based on data integration.

Data integration, however, performed without taking into account the hierarchical relationships between the functional classes, exhibits serious inconsistencies due to the violation of the *true path rule*, governing the functional annotations of genes both in the GO and in FunCat taxonomies (The Gene Ontology Consortium, 2000; Ruepp et al., 2004).

A possible approach to the solution of this problem consists in combining independent local predictions at each functional node in order to obtain a set of probabilistic predictions that are consistent with both the topology and the relational constraints underlying the functional ontology (Barutcuoglu et al., 2006). This approach has been recently investigated in a whole genome and whole ontology gene function prediction experiment, which demonstrated that hierarchical multilabel methods can play a crucial role for the improvement of gene function prediction performances (Obozinski et al., 2008). Nevertheless the approach suffers from some drawbacks. First, it is based on the evaluation of mouse data, and recently published systematic studies showed that the quality of gene functional annotations in this organism are lower than the ones available for the model organism *S. cerevisiae* (yeast) (Buza et al., 2008). Second, the paper focuses on the comparison of hierarchical multilabel methods, but it does not take into account the im-

Appearing in *Working Notes of the 2nd International Workshop on Learning from Multi-Label Data*, Haifa, Israel, 2010.

pact of the concurrent use of data integration and hierarchical multilabel methods on the overall classification performances. Moreover, potential improvements could be introduced by applying cost sensitive variants of hierarchical multilabel predictors, able to effectively calibrate the precision/recall trade-off at different levels of the functional ontology.

In this work we propose a new methodological approach for integrating hierarchical multi-label techniques, data fusion, and cost-sensitive methods. We investigate the impact of these techniques, and their possible synergic effects, on the gene function prediction performance with the yeast model organism. More specifically, we integrate previously studied data fusion methods (Re & Valentini, 2010) and hierarchical multi-label cost-sensitive algorithms (Cesa-Bianchi & Valentini, 2010) to perform a genome and ontology-wide classification of genes according to the FunCat taxonomy.

In Section 2 we present the proposed methods; in Section 3 we summarize the experimental set-up and in Section 4 we discuss the results of the genome and ontology-wide multi-label hierarchical classification. Section 5 contains the conclusions.

2. Methods

2.1. Basic notation

We represent a gene g with a vector $\boldsymbol{x} \in \mathbb{R}^d$ having d different features (e.g., presence or absence of interactions with other d genes, or gene expression levels in different d conditions). A gene g is assigned to one or more functional classes in the set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$ structured according to a FunCat tree T [1]. The assignments are coded through a vector of multilabels $\mathbf{v} = (v_1, v_2, \ldots, v_m) \in \{0,1\}^m$, where g belongs to class ω_i if and only if $v_i = 1$.

In the FunCat tree T, nodes correspond to classes, and edges to relationships between classes. We denote with i the node corresponding to class ω_i. We represent by child(i) the set of nodes that are children of i, and by par(i) the set of parents of i. Moreover, $v_{\text{par}(i)}$ denotes the label of the parent class of i.

The multilabel of a gene g is built starting from the set of the most specific classes occurring in the gene's FunCat annotation; we add to them all the nodes on paths from these most specific nodes to the root. This "transitive closure" operation ensures that the resulting multilabel satisfies the *true path rule*, by which if g belongs to a class/node i, then it also belongs to par(i).

[1] The root of T is a dummy class ω_0, which every gene belongs to, that we added to facilitate the processing

2.2. Data fusion techniques

The data integration is performed locally at each node/class of the FunCat taxonomy. We consider two techniques: ensemble (weighted voting) and kernel fusion.

Let $V_i \in \{0,1\}$ be the random variable that models the labeling of a gene g for the class $\omega_i \in \Omega$. Given L different sources of biomolecular data D_t, for $t = 1, \ldots, L$, we train node classifiers $c_{t,i}$ on the data set D_t, one for each class $\omega_i, 1 \le i \le m$. Let $\widehat{p}_{t,i}(g)$ be the classifier's estimate of the probability that g belongs to ω_i.

A simple way to integrate L different data sources is via the weighed linear combination rule (Kittler et al., 1998). The resulting ensemble estimates the probability that a given gene g belongs to class ω_i by a convex combination of the probabilities estimated by each base learner trained on a different "view" of the data:

$$\hat{P}(V_i = 1 \mid g) = \frac{1}{\sum_{s=1}^{L} F_s} \sum_{t=1}^{L} F_t \, \widehat{p}_{t,i}(g) \qquad (1)$$

where F_t is the F-measure assessed on the training data for the t-th base learner. The choice of the F-measure instead of the accuracy is motivated by the fact that gene classes are largely unbalanced (there are fewer positive examples than negative ones). Given a gene g, the decision $\widehat{y}_i$ of the ensemble about the class ω_i is taken using estimates (1),

$$\widehat{y}_i = \begin{cases} 1, & \text{if } \hat{P}(V_i = 1 \mid g) > \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

where output 1 corresponds to assigning class ω_i to g.

Another popular method to combine different sources of data is kernel fusion (Lanckriet et al., 2004). Kernel fusion (KF) for data integration is based on the closure property of kernels with respect to the sum or other algebraic operators. Given a pair of genes g and g', their corresponding pairs of feature vectors $\boldsymbol{x}_t, \boldsymbol{x}'_t \in D_t$, we implement a kernel averaging function $K_{\text{ave}}(g, g')$ by simply averaging the output of kernel functions $K_1, \ldots, K_L$ specific to each data set,

$$K_{\text{ave}}(g, g') = \frac{1}{L} \sum_{t=1}^{L} K_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \; . \qquad (3)$$

In our experiments we integrated the different data sets by simply summing their normalized kernel matrices. Then we trained the SVM using the resulting matrix. In this case we also use probabilistic

SVMs (Lin et al., 2007) in order to obtain estimates of the posterior probability $\mathbb{P}(V_i = 1 \mid g)$ that a given gene g belongs to class ω_i.

2.3. Hierarchical multi-label cost-sensitive ensemble methods

Recall that $\widehat{p}_i(g)$ is the estimate of the probability that gene g belongs to class ω_i, for $i = 1, \ldots, m$. In this subsection we describe a number of ensemble methods that, given $\widehat{p}_1(g), \ldots, \widehat{p}_m(g)$, derive a multilabel assignment $\widehat{\boldsymbol{y}} = (\widehat{y}_1, \ldots, \widehat{y}_m) \in \{0,1\}^m$ to the nodes of the taxonomy T. The first ensemble method that we consider is HBAYES —see, e.g., (Cesa-Bianchi & Valentini, 2010). This method assumes that, given a gene g, the distribution of the labels $\boldsymbol{V} = (V_1, \ldots, V_m)$ is $\mathbb{P}\big(\boldsymbol{V} = \boldsymbol{v}\big) = \prod_{i=1}^m p_i(g)$ for all $\boldsymbol{v} \in \{0,1\}^m$, where

$$p_i(g) = \mathbb{P}\big(V_i = v_i \mid V_{\mathrm{par}(i)} = 1,\, g\big) \ .$$

In order to the true path rule, we impose that $\mathbb{P}\left(V_i = 1 \mid V_{\mathrm{par}(i)} = 0,\, g\right) = 0$ for all nodes i and all genes g. This implies that the base learner at node i is only trained on the subset of the training set including all examples such that $v_{\mathrm{par}(i)} = 1$.

In the evaluation phase, HBAYES predicts the Bayes-optimal multilabel $\widehat{\boldsymbol{y}} \in \{0,1\}^m$ for a gene g based on the estimates $\widehat{p}_i(g)$ for $i = 1, \ldots, m$. Namely,

$$\widehat{\boldsymbol{y}} = \operatorname*{argmin}_{\boldsymbol{y} \in \{0,1\}^m} \mathbb{E}\big[\ell_H(\boldsymbol{y}, \boldsymbol{V}) \mid g\big] \tag{4}$$

where the expectation is w.r.t. the distribution of $\boldsymbol{V}$. Here $\ell_H(\boldsymbol{y}, \boldsymbol{V})$ denotes the H-loss (Cesa-Bianchi et al., 2006), measuring a notion of discrepancy between the multilabels $\boldsymbol{y}$ and $\boldsymbol{V}$. The main intuition behind the H-loss is simple: *if a parent class has been predicted wrongly, then errors in its descendants should not be taken into account*. Given fixed *cost coefficients* $c_1, \ldots, c_m > 0$, $\ell_H(\widehat{\boldsymbol{y}}, \boldsymbol{v})$ is computed as follows: all paths in the taxonomy T from the root down to each leaf are examined and, whenever a node $i \in \{1, \ldots, m\}$ is encountered such that $\widehat{y}_i \neq v_i$, then c_i is added to the loss, while all the other loss contributions from the subtree rooted at i are discarded. In order to control the sparsity of the multilabels generated by HBAYES in the evaluation phase, we set the base cost coefficients to $c_i = 1/|\mathrm{root}(T)|$, if $i \in \mathrm{root}(T)$, otherwise $c_i = c_j/|\mathrm{child}(j)|$ with $j = \mathrm{par}(i)$. This normalizes the H-loss in the sense that the maximal H-loss contribution of all nodes in a subtree excluding its root equals that of its root. Note that the cost coefficients do not enter in the calculation of the empirical performances reported in Section 3.

Now, a finer control is achieved by introducing $c_i^- = c_i^+ = c_i/2$. These are the costs respectively associated to a false negative (FN) and a false positive (FP) mistake. Let $\{\, A \,\}$ be the indicator function of event A. Given g and the estimates $\widehat{p}_i = \widehat{p}_i(g)$ for $i = 1, \ldots, m$, the HBAYES prediction rule can be formulated as follows:

HBAYES prediction rule

Initially, set the labels of each node i to

$$\widehat{y}_i = \operatorname*{argmin}_{y \in \{0,1\}} \Big(c_i^- \widehat{p}_i (1-y) + c_i^+ (1-\widehat{p}_i) y + \widehat{p}_i \{y = 1\} \sum_{j \in \mathrm{child}(i)} H_j(\widehat{\boldsymbol{y}}) \Big) \tag{5}$$

where

$$H_j(\widehat{\boldsymbol{y}}) = c_j^- \widehat{p}_j (1-\widehat{y}_j) + c_j^+ (1-\widehat{p}_j) \widehat{y}_j + \widehat{p}_j \{\widehat{y}_j = 1\} \sum_{k \in \mathrm{child}(j)} H_k(\widehat{\boldsymbol{y}})$$

is recursively defined over the nodes j in the subtree rooted at i with each $\widehat{y}_j$ set according to (5).

Then, if $\widehat{y}_i$ is set to zero, set all nodes in the subtree rooted at i to zero as well.

As shown in (Cesa-Bianchi et al., 2006), $\widehat{\boldsymbol{y}}$ can be computed for a given g via a simple bottom-up message-passing procedure whose only parameters are the estimates $\widehat{p}_i$. Unlike standard top-down hierarchical methods —see the description of HTD at the end of this section, each $\widehat{y}_i$ also depends on the classification of its child nodes. In particular, if all child nodes k of i have $\widehat{p}_k$ close to a half, then the Bayes-optimal label of i tends to be 0 irrespective of the value of $\widehat{p}_i$. Vice versa, if i's children all have $\widehat{p}_k$ close to either 0 or 1, then the Bayes-optimal label of i is based on $\widehat{p}_i$ only, ignoring the children —see also (6).

We now introduce a simple cost-sensitive variant, HBAYES-CS, of HBAYES, which is suitable for learning datasets whose multilabels are sparse (i.e., the classes are unbalanced). This variant introduces a parameter α that is used to trade-off the cost of false positive (FP) and false negative (FN) mistakes. We parametrize the relative costs of FP and FN by introducing a factor $\alpha \geq 0$ such that $c_i^- = \alpha c_i^+$ while keeping $c_i^+ + c_i^- = 2c_i$. Then (5) can be rewritten as

$$\widehat{y}_i = 1 \iff \widehat{p}_i \left(2c_i - \sum_{j \in \mathrm{child(i)}} H_j \right) \geq \frac{2c_i}{1+\alpha} \ . \tag{6}$$

This is the rule used by HBAYES-CS in our experiments.

Given a set of trained base learners providing estimates $\widehat{p}_1, \ldots, \widehat{p}_m$, we compare the quality of the multilabels computed by HBAYES-CS with that of HTD-CS. This is a cost-sensitive version of the basic top-down hierarchical ensemble method HTD whose predictions are computed in a top-down fashion (i.e., assigning $\widehat{y}_i$ before the label of any j is the subtree rooted at i) using the rule $\widehat{y}_i = \left\{\widehat{p}_i \geq \frac{1}{2}\right\} \times \{\widehat{y}_{\mathrm{par}(i)} = 1\}$ for $i = 1, \ldots, m$ (we assume that the guessed label $\widehat{y}_0$ of the root of T is always 1). The variant HTD-CS introduces a single cost sensitive parameter $\tau > 0$ which replaces the threshold $\frac{1}{2}$. The resulting rule for HTD-CS is then $\widehat{y}_i = \{\widehat{p}_i \geq \tau\} \times \{\widehat{y}_{\mathrm{par}(i)} = 1\}$.

Note that both methods HBAYES-CS and HTD-CS use the same estimates $\widehat{p}_i$. The only difference is in the way the classifiers are defined in terms of these estimates.

2.4. Integration of Hierarchical multi-label and data fusion methods

The hierarchical ensemble methods combine the probabilistic output of the classifiers associated to each node of the tree. Hence, it is quite straightforward to replace the classifiers trained on single sources of data with ensembles of classifiers trained on multiple sources of data, or with SVMs trained on kernel matrices obtained by summing kernel matrices specific for each data set. To this end we can apply a two-step strategy:

1. Train a set of classifiers that estimate $\mathbb{P}(V_i = 1 \mid g)$ for each node $i = 1, \ldots, m$ of the FunCat taxonomy. Each classifier is an ensemble of base learners, or a SVM trained with multiple sources of data by kernel fusion methods (see Section 2.2).
2. Combine the predictions at each node to obtain the multi-label predictions according to the hierarchical multi-labels methods described in Section 2.3.

The resulting hierarchical multi-label predictions respect the "true path rule", and implement a local combination of multiple sources of biomolecular data at each node of the FunCat tree.

3. Experimental set-up

3.1. Genomic data sets

We integrated six different sources of yeast biomolecular data, previously used for single-source ontology-wide gene function prediction (Cesa-Bianchi & Valentini, 2010). The data sets include two types of protein domain data (PFAM BINARY and PFAM LOGE) downloaded from the Pfam data base (Finn et al., 2008); gene expression measures (EXPR) relative to different conditions (Gasch et al., 2000); protein-protein interaction data (BIOGRID) downloaded from the *BioGRID* data base (Stark et al., 2006) and from the *STRING* data base (STRING) (vonMering et al., 2003); SEQ. SIM. pairwise similarity data that contain log-E values obtained by Smith and Waterman pairwise alignments between all pairs of yeast sequences.

We considered only yeast genes common to all data sets, and in order to get a not too small set of positive examples for training, for each data set we selected only the FunCat-annotated genes and the classes with at least 20 positive examples. This selection process yielded 1901 yeast genes annotated to 168 FunCat classes distributed across 16 trees and 5 hierarchical levels. We added a "dummy" root node to obtain a tree from the overall FunCat forest (Fig. 1). We adopted the following strategy to select negative examples: at each FunCat node the negatives are the genes that are not annotated at the corresponding class, but are annotated at the parent class/node.

3.2. Experimental tasks and performance assessment

We performed several experimental classification tasks at genome and ontology-wide level (i.e., we considered all genes and all the 168 classes of the hierarchically structured multi-label classification problem):
(a) Comparison of "single-source" and data fusion techniques (kernel fusion and weighted voting) using both FLAT and hierarchical methods (HTD and HBAYES);
(b) Assessment of the improvements achievable by: (i) multi-label hierarchical methods vs flat methods; (ii) cost-sensitive vs cost-insensitive strategies; (iii) synergic enhancements due to the concurrent application of multi-label hierarchical methods, cost-sensitive and data fusion techniques;
(c) Analysis of the precision-recall characteristics of the compared methods.
Note that for FLAT ensembles we mean a set of base learners each one predicting a single functional class, without any combination of the predictions that takes into account the hierarchical structure of the classes.

We used linear SVMs with probabilistic output (Lin et al., 2007) as base learners and, following the experimental set-up proposed by Lewis et al. (2006), we did not perform model selection at this level (we simply set the regularization parameter C to 10). To assess the generalization capabilities of the ensem-

ble, we adopted "external" 5-fold cross validation techniques, while to select the threshold value τ for HTD-CS ensembles and the α value for for HBAYES-CS ensembles we applied "internal" 3-fold cross-validation.

In the context of ontology-wide gene function prediction problems, where negative examples are usually a lot more than positives, accuracy is not a reliable measure to assess the classification performance. For this reason we adopted the classical F-score to take into account the unbalance of FunCat classes. Moreover, in order to better capture the hierarchical and sparse nature of the gene function prediction problem we also applied the *hierarchical F-measure*: this measure is based on the estimation of how much the predicted classification paths correspond to the correct paths, and expresses in a synthetic way the effectiveness of the structured hierarchical prediction (Verspoor et al., 2006). More precisely, for a given gene or gene product x consider the subtree $G(x) \subset T$ of the predicted classes and the subtree $C(x)$ of the correct classes associated to x. For a leaf $f \in G(x)$ and $c \in C(x)$, let be $\uparrow f$ and $\uparrow c$ the set of their ancestors that belong, respectively, to $G(x)$ and $C(x)$. The hierarchical precision (HP) and hierarchical recall (HR) (Verspoor et al., 2006) are defined as follows:

$$HP = \frac{1}{|\ell(G(x))|} \sum_{f \in \ell(G(x))} \frac{|C(x) \cap \uparrow f|}{|\uparrow f|}$$

$$HR = \frac{1}{|\ell(C(x))|} \sum_{c \in \ell(C(x))} \frac{|\uparrow c \cap G(x)|}{|\uparrow c|}$$

where $\ell(\cdot)$ is the set of leaves of a tree. The *hierarchical F-measure* is the harmonic mean of the hierarchical precision and recall.

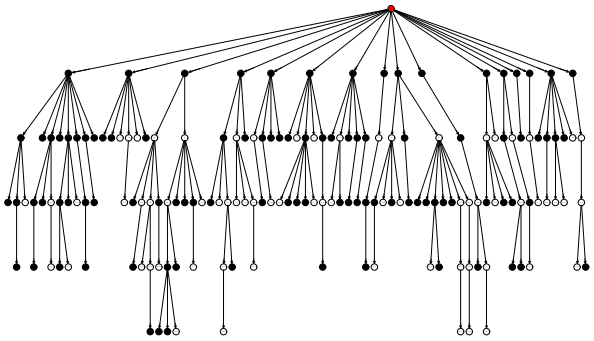
4. Results

4.1. Impact of data fusion on flat and hierarchical methods

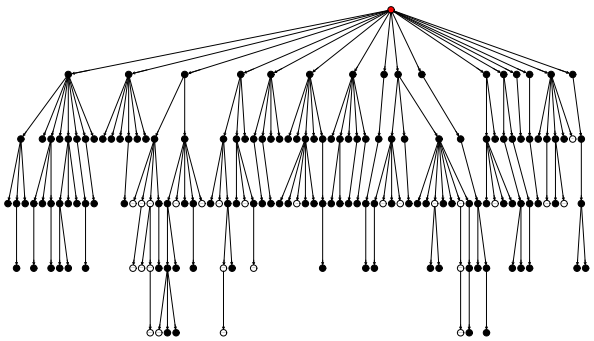
Table 1 summarizes the results of the comparison of single-source and data integration approaches to both flat and hierarchical ensembles. Data fusion techniques improve average per class F-score across classes in FLAT ensembles (first column of Table 1). These results confirm and extend to the entire FunCat ontology previous results limited only to the most general first level classes of the taxonomy (Re & Valentini, 2010). Looking at Fig. 1 (a), we can observe that the increment in performances due to the application of heterogeneous data integration methods is not limited to a specific level of the functional ontology, but spans all the 5 levels of the FunCat tree.

Table 1. Average per-class F scores with FLAT, HTD, HTD-CS, HB (HBAYES) and HB-CS (HBAYES-CS) ensembles, using single sources and multi-source (data fusion) techniques.

METHODS	FLAT	HTD	HTD-CS	HB	HB-CS
SINGLE-SOURCE					
BIOGRID	0.2643	0.3759	0.4160	0.3385	0.4183
STRING	0.2203	0.2677	0.3135	0.2138	0.3007
PFAM BINARY	0.1756	0.2003	0.2482	0.1468	0.2395
PFAM LOGE	0.2044	0.1567	0.2541	0.0997	0.2500
EXPR.	0.1884	0.2506	0.2889	0.2006	0.2781
SEQ. SIM.	0.1870	0.2532	0.2899	0.2017	0.2825
MULTI-SOURCE (DATA FUSION)					
KERNEL FUSION	0.3220	0.5401	0.5492	0.5181	0.5505
WEIGH. VOTING	0.2754	0.2792	0.3974	0.1491	0.3532



(a)



(b)

Figure 1. FunCat trees representing the comparison between F-scores achieved with data integration (KF) vs the best single-source classifiers trained on BIOGRID data. Black nodes depict functional classes for which KF achieves better F-scores. (a) FLAT, (b) HBAYES-CS ensembles.

Multi-label hierarchical methods show similar results (columns HTD, HTD-CS HBAYES and HBAYES-CS of Table 1). Note that Kernel Fusion largely improves on results achieved with any "single-source" ensemble methods, while Weighted Voting results are worse than those of the best single-source (BIOGRID) when

Table 2. Wilcoxon signed-ranks test results to evaluate the statistical significance of the improvement of data fusion techniques w.r.t. single data sources achieved with cost-sensitive multi-label hierarchical methods (HBAYES-CS and HTD-CS). Results in boldface are in favour of ensembles using single data sources.

HBAYES-CS, $\alpha = 2$						
	BIOGRID	STRING	PFAM BIN.	PFAM LOGE	EXPR.	SEQ. SIM.
KERNEL FUSION	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
WEIGHTED VOTING	$\mathbf{2.3 \times 10^{-4}}$	5.6×10^{-07}	2.2×10^{-16}	9.1×10^{-15}	1.3×10^{-15}	3.8×10^{-13}
HTD-CS, $\tau = 0.4$						
	BIOGRID	STRING	PFAM BIN.	PFAM LOGE	EXPR.	SEQ. SIM.
KERNEL FUSION	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
WEIGHTED VOTING	$\mathbf{9.5 \times 10^{-2}}$	6.9×10^{-12}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}

hierarchical ensemble methods are applied (with FLAT ensembles Weighted Voting improves on BIOGRID). These results seem to partially contradict previous ones published in Re & Valentini (2010), but note that in that work only the most general classes at the first level of the FunCat hierarchy were classified, and no hierarchical methods were applied.

The improvements achieved by data integration techniques are statistically significant according to the Wilcoxon test (Table 2). With both HBAYES-CS and HTD-CS hierarchical ensembles, Kernel Fusion performances are significantly better than any single-source approach ($p\text{-}value = 2.2 \times 10^{-16}$). This is true also for Weighted Voting except for the BIOGRID data where results are in favour of this single-source data with both HTD-CS ($p\text{-}value = 9.5 \times 10^{-2}$) and HBAYES-CS ensembles ($p\text{-}value = 2.3 \times 10^{-4}$).

Focusing on Kernel Fusion, Fig. 1 depicts the classes (black nodes) where KF achieves better results than the best single-source data set (BIOGRID). It is worth noting that there is a synergy between KF and hierarchical methods, because the number of black nodes is significantly larger in HBAYES-CS ensembles (Fig. 1 b) w.r.t FLAT methods (Fig. 1 a). It is well-known that hierarchical multi-label ensembles largely outperform FLAT approaches (Guan et al., 2008; Obozinski et al., 2008), but these results show that data fusion techniques can further improve performances w.r.t. FLAT methods.

4.2. Analysis of the synergy between hierarchical multi-label methods, cost sensitive and data fusion techniques

Hierarchical F-score results confirm the results of Section 4.1: data fusion and in particular Kernel Fusion improves performances of ensemble methods. In particular, we obtain a marked improvement with hierarchical ensemble methods (Table 3).

According to previous works (Valentini & Re, 2009; Cesa-Bianchi & Valentini, 2010), cost-sensitive approaches boost predictions of hierarchical methods when single-sources of data are used to train the base learners. These results are confirmed also when cost-sensitive methods (HBAYES-CS and HTD-CS) are integrated with data fusion techniques, showing a synergy between multi-label hierarchical, data fusion (in particular kernel fusion), and cost-sensitive approaches (Table 3). The improvements of per-class F-scores achieved by HBAYES-CS and HTD-CS are statistically significant at 0.005 significance level (Wilcoxon test) w.r.t. their "vanilla" counterparts and FLAT methods. No significant difference can be detected between HBAYES-CS and HTD-CS (Table 4). It is worth noting that other approaches for learning unbalanced classes, i.e., undersampling techniques or cost-sensitive SVMs (Morik et al., 1999), can be applied to predict gene functions. They represent local methods that could in principle be combined with the global cost-sensitive approach of HBAYES-CS to further improve prediction performances.

Table 3. Comparison of hierarchical F-score, precision (Prec.) and recall (Rec.) among different ensemble methods using the best source of biomolecular data (BIOGRID), Kernel Fusion (KF), and Weighted Voting (WVOTE) data integration techniques.

METHODS	F-SCORE	PREC.	REC.
BIOGRID:			
FLAT	0.1893	0.1253	0.5801
HTD	0.4311	0.5901	0.3827
HTD-CS	0.4732	0.5645	0.4650
HBAYES	0.3776	0.5404	0.3236
HBAYES-CS	0.4738	0.5654	0.4639
KF:			
FLAT	0.2052	0.1293	0.7026
HTD	0.5800	0.7051	0.5560
HTD-CS	0.6091	0.6745	0.6156
HBAYES	0.5512	0.6915	0.5086
HBAYES-CS	0.6073	0.6759	0.6126
WVOTE:			
FLAT	0.1851	0.1252	0.5265
HTD	0.3183	0.4673	0.2718
HTD-CS	0.4477	0.5838	0.4148
HBAYES	0.1729	0.2639	0.1445
HBAYES-CS	0.4053	0.5437	0.3691

Table 4. Wilcoxon signed-ranks test results to evaluate the statistical significance of the improvement of cost-sensitive w.r.t non cost-sensitive multi-label hierarchical methods. Data integration method: Kernel Fusion.

	FLAT	HTD	HTD-CS	HBAYES	HBAYES-CS
HBAYES-CS ($\alpha = 2$)	2.2×10^{-16}	5.9×10^{-04}	1.6×10^{-01}	1.1×10^{-14}	–
HTD-CS ($\tau = 0.4$)	2.2×10^{-16}	2.9×10^{-03}	–	2.8×10^{-13}	8.3×10^{-01}

Per-level analysis of the F-score in HBAYES-CS and HTD-CS ensembles, shows a certain degradation of performance w.r.t. the depth of nodes (Table 5), but this degradation is largely lower when data fusion is applied. Indeed in Cesa-Bianchi & Valentini (2010) the per-level F-score achieved by HBAYES-CS and HTD-CS when a single source is used continuously decreases from the top to the bottom level, and it is halved at level 5 w.r.t. to the first level, while in our experiments with Kernel Fusion the average F-score at level 2, 3 and 4 is comparable, and the decrement at level 5 w.r.t. level 1 is reduced at about 15% (Table 5).

In conclusion, the synergic effects of hierarchical multi-label ensembles, cost-sensitive and data fusion techniques significantly improve performances of gene function prediction. Moreover these enhancements permit to obtain better and more homogeneous results at each level of the hierarchy. This is of paramount importance, because more specific annotations are more informative and can get more biological insights into the functions of genes.

4.3. Analysis of the precision/recall characteristics of hierarchical multi-label methods

Hierarchical precision/recall results using the best single-source of data and data fusion techniques show that FLAT methods achieve the best recall and HTD the best precision (except for Weighted Voting where HBAYES-CS and HTD-CS obtain the best precision, Table 3); HBAYES-CS and HTD-CS are in between, achieving good "intermediate" results for both precision and recall, thus resulting in the best F-score.

Note that hierarchical precision of FLAT methods is too low to be useful in practice, and precision of HBAYES-CS and HTD-CS is quite close to that of HTD ensembles, that suffer from a significantly lower recall (Table 3).

Interestingly enough, while the overall hierarchical precision and recall between HBAYES-CS and HTD-CS is quite similar (Table 3), the average precision at the low levels of the FunCat taxonomy is higher in HBAYES-CS (Table 5). Fig. 2 shows that the black nodes representing FunCat classes for which HBAYES-CS improves precision are concentrated at the middle and lower levels of the hierarchy. This is of paramount importance in real applications, when we need to reduce the costs of the biological validation of new gene functions discovered through computational methods.

Table 5. Per level average Precision (P), Recall (R), Specificity (S), F-score (F) and Accuracy (A) across the five levels of the FunCat taxonomy in HBAYES-CS and HTD-CS ensembles using Kernel Fusion data integration. Level 1 is the top level, level 5 the bottom.

HBAYES-CS, $\alpha = 2$					
LEVEL	P	R	S	F	A
1	0.7071	0.5399	0.9523	0.6025	0.9052
2	0.6793	0.4785	0.9817	0.5447	0.9570
3	0.6452	0.5059	0.9893	0.5514	0.9755
4	0.5874	0.5318	0.9875	0.5428	0.9759
5	0.5741	0.4704	0.9942	0.5048	0.9871
HTD-CS, $\tau = 0.4$					
LEVEL	P	R	S	F	A
1	0.7104	0.5399	0.9525	0.6029	0.9051
2	0.6638	0.4832	0.9810	0.5440	0.9565
3	0.6257	0.5187	0.9882	0.5528	0.9747
4	0.5461	0.5441	0.9865	0.5364	0.9752
5	0.5284	0.4823	0.9933	0.4924	0.9863

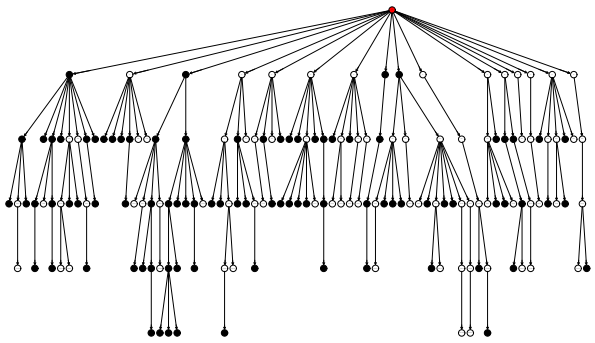


Figure 2. Ontology-wide FunCat tree plot highlighting the nodes at which the precision of the HBAYES-CS is larger than the one obtained by HTD-CS, using Kernel Fusion to integrate multiple sources of data.

Another advantage of HBAYES-CS is represented by the fact that its precision/recall characteristics can be tuned via a single global parameter, the cost factor $\alpha = c_i^- / c_i^+$: by incrementing α we introduce progressively lower costs for positive predictions, thus resulting in an increment of the recall (at the expenses of a possibly lower precision, results not shown). More-

over, by setting the α parameter at each node to the ratio of negative and positive examples for the corresponding class (Cesa-Bianchi & Valentini, 2010), we can reach results comparable with those obtained by internal cross-validation of the global α parameter, thus avoiding a certain computational burden (results not shown).

5. Conclusions

In this work we demonstrated that the combined use of heterogeneous data integration methods performed locally, followed by a global probabilistic reconciliation of the predictions produced at each node is more effective than the hierarchical combination of classifiers trained using single data-sets. These results are strengthened when a cost-sensitive strategy is applied to deal with the unbalance between positive and negative examples.

The results confirmed also that the increment in performances due to data integration methods is not limited to specific levels of the FunCat taxonomy and that, among the best performing hierarchical multi-label methods, the HBAYES-CS ensemble is able to more effectively preserve the precision across the ontology levels (in particular near to the leaves) than the HTD-CS ensemble method.

The synergy between heterogeneous data integration, hierarchical multi-label and cost-sensitive approaches is the key to drive bio-molecular experiments aimed at the discovery of previously unannotated gene functions. Considering the performance in terms of both precision and recall, no significant difference can be detected between the two proposed hierarchical cost-sensitive ensembles, HBAYES-CS and HTD-CS. Nevertheless, among the compared algorithms, HBAYES-CS, integrated with data fusion methods, is the best choice to ensure the high precision level required in large scale gene function prediction projects.

References

Barutcuoglu, Z., Schapire, R.E., and Troyanskaya, O.G. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

Buza, T.J. et al. Gene ontology annotation quality analysis in model eukaryotes. *Nucleic Acids Research*, 36, 2008.

Cesa-Bianchi, N. and Valentini, G. Hierarchical cost-sensitive algorithms for genome-wide gene function prediction. *JMLR, Machine Learning in Systems Biology*, 8:14–29, 2010.

Cesa-Bianchi, N., Gentile, C., and Zaniboni, L. Incremental algorithms for hierarchical classification. *JMLR*, 7: 31–54, 2006.

Finn, R.D. et al. The Pfam protein families database. *Nucleic Acids Research*, 36:D281–D288, 2008.

Gasch, P. et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol.Biol.Cell*, 11:4241–4257, 2000.

Guan, Y. et al. Predicting gene function in a hierarchical context with an ensemble of classifiers. *Genome Biology*, 9(S2), 2008.

Karaoz, U. et al. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl Acad. Sci. USA*, 101:2888–2893, 2004.

Kittler, J. et al. On combining classifiers. *IEEE Trans. on PAMI*, 20(3):226–239, 1998.

Lanckriet, G.R. et al. A statistical framework for genomic data fusion. *Bioinformatics*, 20:2626–2635, 2004.

Lewis, D.P., Jebara, T., and Noble, W.S. Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. *Bioinformatics*, 22(22):2753–2760, 2006.

Lin, H.T., Lin, C.J., and Weng, R.C. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68:267–276, 2007.

Morik, K., Brockhausen, P., and Joachims, T. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. In *Proc of 16th ICML*, 1999.

Obozinski, G. et al. Consistent probabilistic outputs for protein function prediction. *Genome Biology*, 9 Suppl. 1:S6, 2008.

Pavlidis, P. et al. Learning gene functional classifications from multiple data types. *J Comput Biol*, 9(2), 2002.

Re, M. and Valentini, G. Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction. *JMLR, Machine Learning in Systems Biology*, 8:98–111, 2010.

Ruepp, A. et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research*, 32(18), 2004.

Stark, C. et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, 34:D535–D539, 2006.

The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genet.*, 25:25–29, 2000.

Valentini, G. and Re, M. Weighted True Path Rule: a multilabel hierarchical algorithm for gene function prediction. In *MLD-ECML 2009, 1st Int. Workshop on learning from Multi-Label Data*, pp. 133–146, 2009.

Verspoor, K. et al. A categorization approach to automated ontological function annotation. *Protein Science*, 15:1544–1549, 2006.

vonMering, C. et al. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31:258–261, 2003.