# Gene-Disease Prioritization Through Cost-Sensitive Graph-Based Methodologies

Marco Frasca(✉) and Simone Bassis

Department of Computer Science, University of Milano,
Via Comelico 39/41, 20135 Milano, Italy
{frasca,bassis}@di.unimi.it

**Abstract.** Finding genes associated with human genetic disorders is one of the most challenging problems in bio-medicine. In this context, to guide researchers in detecting the most reliable candidate causative-genes for the disease of interest, gene prioritization methods represent a necessary support to automatically rank genes according to their involvement in the disease under study. This problem is characterized by highly unbalanced classes (few causative and much more non-causative genes) and requires the adoption of cost-sensitive techniques to achieve reliable solutions. In this work we propose a network-based methodology for disease-gene prioritization designed to expressly cope with the data imbalance. Its validation over a benchmark composed of 708 selected medical subject headings (MeSH) diseases, shows that our approach is competitive with state-of-art methodologies, and its reduced time complexity makes its application feasible on large-size datasets.

**Keywords:** Gene-disease prioritization · Graph-based node ranking · Cost-sensitive learning

## 1 Introduction

Linkage studies for determining relevant genes for specific human diseases can point to a genomic region containing hundreds of genes, while the high-throughput sequencing approach will often identify a great number of non-synonymous genetic variants. Although the detection of potentially deleterious variants can be easily automated, this can often result in the identification of thousands candidate disease genes. Since the experimental verification of an individual gene can be both difficult and time consuming, an efficient way to reduce the validation cost is to narrow down the large list of candidate genes to a small and manageable set of promising genes; a process called *gene prioritization* (GP).

As manual examination of biological databases in order to select the most promising causative genes for the disease of interest has been only partially successful, since the selection is based solely on the subjective impressions of the researcher and genetic disorders often involve several primarily responsible genes, various computational GP methods have been proposed for this purpose.


F. Ortuño and I. Rojas (Eds.): IWBBIO 2016, LNBI 9656, pp. 739–751, 2016.
DOI: 10.1007/978-3-319-31744-1_64

Earlier works [1] investigated gene-diseases associations based on *gene expression profiles* or *genome wide association studies* (GWAS). Genome-wide association studies identify genes involved in human disease by searching the genome for small variations, called *single nucleotide polymorphisms* (SNPs), that occur more frequently in people with a particular disease than in healthy people. Each study can look at hundreds or thousands of SNPs at the same time. However, this approach tends to produce many false positive results, and the experimental validation of these candidate genes, for instance through resequencing, pathway or expression analysis, is still expensive and time consuming [2].

For these reasons other GP approaches have been investigated, such as *guilt-by-association* (GBA), in which candidate disease genes are ranked by exploiting the assumption that similar genes tend to share similar diseases [3]. The input of these methods is represented by gene networks, in which nodes represent genes and connections encode precomputed functional relationships among genes, such as common functional annotations (e.g. Gene Ontology annotations [4]), transcriptional co-expression regulation, direct molecular interactions [5]. In this context, many approaches have been adopted to compute the GP ranking, ranging from protein-protein interaction network analysis and semi-supervised graph partitioning [5], to flow propagation [6], and random walks [7].

To improve the accuracy of GP methods, recent studies have investigated the advantage of integrating multiple data sources, including expression profiles, SNP genotype data, expression quantitative trait loci, functional profiles, and network-based sources, such as gene-chemical networks, protein complexes and genetics/physical interactions [8]. A general approach in data source integration ranks each candidate gene according to each individual data source using various metrics, and then combine ranks from all data sources by using order statistics to obtain an overall rank [3]. For network-based integration approaches, a consensus network is constructed by combining the structure and the characteristics of each network, through different network integrating strategies [9]. The consensus network tends to provide better signal-to-noise ratio and complementary information about genes, thus leading to an improvement in prediction accuracy in most of cases [9,10].

Apart from the disadvantages and the benefits discussed above for each different approach, the main drawback shared by the above-mentioned GP methods is that they completely neglect the class imbalance problem characterizing GP: there are much fewer causative genes (the positive instances) than non-causative ones (the negative instances). For instance, around 40 % (10/09/15 update) of known genetic diseases in the OMIM (Online Mendelian Inheritance in Man) database have still fewer or almost none established gene-disease associations [11]. Computational methodologies usually suffer from a drastic performance deterioration in case of imbalance classes, since algorithms tend more to focus on the classification of major class samples while ignoring or misclassifying minority class samples [12]. Unfortunately, in our context the minority class carries almost all the information we have about the disease under study, and this makes necessary the adoption of specifically designed imbalance-aware

machine learning algorithms, often referred to as *cost-sensitive*. For instance, cost-sensitive techniques obtained successful result in similar contexts, e.g. in the protein function prediction [13,14].

Here we propose a novel network-based approach for detecting disease-gene association which aims at coping with the label imbalance by 'transforming' the input network so as to effectively represent the label imbalance, and by applying cost-sensitive methodologies on the obtained network representation. In particular, our procedure can be summarized as follows: (1) by following the approach proposed in [15], the input network is projected onto a bidimensional space, where each labeled input node corresponds to a labeled point whose coordinates depend on its positive and negative neighborhood in the input network, respectively; (2) the obtained couple of coordinates/features for each point are given in input to a cost-sensitive family of regressors to learn an cost-sensitive model to rank the unlabeled nodes. The node projection at Step 1 embeds the imbalance between positive and negative genes at each neighborhood in the corresponding point position. Moreover, working with just two features makes the Step 2 of our procedure very fast, thus allowing our method to efficiently handle large data sets. Finally, the method in general enough to include strategies for integrating heterogeneous network sources in a dedicated preprocessing step, so as to exploit the benefit of working with more reliable and informative networks. We experimentally validated our method on a public benchmark data set for GP, including almost nine thousands of human genes and around seven hundreds diseases collected from the Medical Subject Headings database[1].

The paper is organized as follows: in Sect. 2 we formalize the problem, while Sect. 3 is devoted to describe both the gene networks and the network integration techniques adopted in the benchmark experimental setting. In Sect. 4 we introduce our proposed two-step procedure; then in Sect. 5 we check its effectiveness by comparing its performance with state-of-the-art methodologies. Finally, Sect. 6 concludes the paper.

## 2 Problem Setting

The disease-gene prioritization problem can be seen as a semi-supervised bipartite ranking problem on undirected graphs [16]. Specifically, a gene network can be represented through an undirected weighted graph $G = (V, \boldsymbol{W})$, where $V = \{1, 2, \ldots, n\}$ is the set of vertices corresponding to genes, and $\boldsymbol{W}$ is the $n \times n$ weight matrix, where each element $W_{ij} \in [0, 1]$ represents some notion of functional similarity between vertices $i$ and $j$. Vertices in $V$ can be partitioned into two subsets: $S \subset V$ containing instances labeled according to a specific MeSH subject heading, and its complement $U = V \setminus S$, including unlabeled instances and therefore representing the object of our inference. As for the former, the set of positive/negative instances are denoted respectively with $S_+$ and $S_-$.

The task we are called to solve consists in learning a ranking function $\phi : U \to \mathbb{R}$ that assigns values to future positive instances higher than to negative ones,

[1] http://www.nlm.nih.gov/mesh.

ranking therefore the former higher than the latter. From this standpoint, GP is cast as a semi-supervised learning problem on graphs, since gene ranking can be inferred by exploiting both labeled and unlabeled nodes (genes) and the connections among them.

To make the problem even harder, the family of graphs under investigation is subjected to a strong imbalance between negative and positive instances, presenting a strong disproportion in favor of negative labeled nodes.

## 3 Materials

The input connection matrix $\boldsymbol{W}$ represents a complex set of interactions or similarities between genes and/or their products (such as proteins), obtained as integration of several heterogeneous data sources. We adopt the benchmark experimental setting proposed in [9], which is composed of nine human gene networks covering 8449 genes, and describing functional interactions, transcriptional co-expression/regulation and localization, gene expression profiles,

**Table 1.** Gene networks used in experimental campaign.

| | |
|---|---|
| *finet* | *Functional interaction network* – A network covering 8441 selected proteins and containing protein-protein interactions inferred by a Naive Bayes classifier [18]. |
| *hnnet* | *Human net* – Functional gene network integrating 21 large-scale genomics and proteomics datasets from four species [19], spanning diverse distinct lines of evidence. |
| *cmnet* | *Cancer module network* – Gene-gene network composed of 8849 genes, where two genes are connected if they share at least one of the 263 biological and clinical conditions considered in [20], collecting expression profiles in different tumors. |
| *gcnet* | *Gene chemical network* – A network of 7649 genes constructed starting from the genes-chemicals interactions available at the CTD database. |
| *dbnet* | *BioGRID database network* – A protein-protein interaction network of 8449 proteins based upon direct physical and genetic interactions obtained from BioGRID [21] |
| *bgnet* | *BioGRID projected network* – Network obtained by: (i) constructing a bipartite graph exploiting interactions between genes available in BioGRID; and (ii) inserting an edge between two genes if they share at least one neighbor in the bipartite graph. |
| *bpnet, mfnet, ccnet* | *Semantic similarity-based networks* – Three networks obtained by considering the Gene Ontology terms [4] in the three branches (biological process, molecular function, and cellular component). Each connection weight is the maximum Rensik semantic similarity between all the terms for which the two genes are GO annotated. |

genes-chemicals relationships, protein-protein physical and genetic interactions, and GO semantic similarity (see Table 1). The database also provides the associations of such genes with 708 selected MeSH (Medical Subject Headings) diseases, downloaded from the CTD database [17]. The selected MeSH disease terms include between 5 and 200 causative genes.

### 3.1 Network Integration

Since the various networks have different number of genes, before their combination we extend them to the union of genes in the single networks, by filling each network with zeros in the corresponding missing rows/columns. As done in [9], in a pre-processing step we delete smaller edges so as to remove too small (and putative noisy) similarities, and ensure at least one neighbor for each node.

As integration scheme we adopt the *unweighted integration* of single networks, which performed better among the unweighted schemes proposed in [9]. It is the simple average of the $m$ available network adjacency matrices, i.e. $\boldsymbol{W}^* = \sum_{d=1}^{m} \boldsymbol{W}^{(d)}/m$. Finally, we apply to $\boldsymbol{W}^*$ the Laplacian normalizazion $\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}^*\boldsymbol{D}^{-\frac{1}{2}}$, where $\boldsymbol{D}$ is a diagonal matrix $D_{ii}|_{i=1}^{n}$, with $D_{ii} = \sum_j \boldsymbol{W}^*_{ij}$.

We performed our experimentations on two networks: the first, called Net6 hereinafter, was obtained by integrating the six gene networks with the exclusion of the semantic similarity-based ones; the latter (Net9) is defined as unweighted integration of all the nine single networks reported in Table 1.

## 4 Methods

We decided to solve the bipartite ranking problem introduced in Sect. 2 in terms of a generalized linear model (GLM) where the response variable, suitably thresholded through the sign function, decrees the membership to either the positive or negative class, while the predictors have been chosen so as to exploit the nodes similarity coded in the weight matrix $\boldsymbol{W}$. In order to keep the computational burden low and to exploit the network topology, we extract from the input network two features[2], as follows: each node $i \in S$ is associated with a point $\Delta_i = (\Delta_i^+, \Delta_i^-)$ in the plane, where

$$\Delta_i^+ = \sum_{j \in S_+} w_{ij}, \qquad \Delta_i^- = \sum_{j \in S_-} w_{ij} \tag{1}$$

Intuitively, the more node $i$ is functionally similar to positive nodes and the higher will be the value of its $\Delta_i^+$ coordinate; analogously for the contribution given by negative nodes to the second coordinate. Remembering the one-to-one correspondence between genes and vertices, with this projection we hope to find a bipartition of nodes in $S$ which concentrates positive nodes mostly toward the rightmost lower region of the first quadrant, and negative ones in the remaining

[2] Actually the number of predictors, including the two-way interaction term (i.e. the product of the two features), is equal to 3.

portion of it. This network projection onto the plane, already adopted in [15], also allows to both avoid the *curse of dimensionality* problem, since the projected space has just two dimensions, and deal with the class imbalance problem, since the projected positive and negative two-dimensional points can be associated with different misclassification costs during the learning of the GLM.

**Table 2.** GLMs adopted in the experimental campaign.

| | |
|---|---|
| LR | *Linear regression model.* Usually disregarded in case of dichotomous categorical dependent variables, mainly to avoid the risk of "impossible predictions" (i.e. results outside of the unit interval), we include it in our analysis in view of both the straightforward interpretability of its coefficients, the not negligible speed-up factor observed when large datasets are given in input to the model, and the groundlessness of the aforementioned risks when interactions terms are included in the model [22]. |
| LogR | *Logit regression model.* Together with *probit* model, it is one of the widely used regression models for binary response variables. Despite the different assumptions the two models make about the error distribution, results tend to be so similar each other that preference for one over the other model tends to vary by discipline. We opted to work with logistic regression (whose link function reads as $g(x) = \log(x/(1-x))$) mainly for the straightforward interpretation of the estimated coefficients. |
| CLogLR | *Cloglog regression model.* While logit and probit are symmetric link functions, the choice of a skewed link function provides a better fit to unbalanced data [23]. Binomial regression model with complementary log-log link function (defined as $g(x) = \log(-\log(1-x))$) is frequently used when the probability of events' occurrence is very small. |
| PR | *Poisson regression model.* As an alternative to an asymmetric link function, the choice of a discrete and skewed distribution for the response variable is often suggested [24]. Poisson regression with the canonical log link function is widely used in case of binary outcome variables to cope with rare events. Indeed, imbalance classification problems represent a typical scenario to apply Poisson regression, since the main assumption which the model relies on that expected value and variance of the response variable coincide, is always (at least approximately) satisfied. |

We adopted the four GLM models, described in Table 2. Within the various cost-sensitive schemes proposed to allow regression models handling imbalance classes [25], one of the most effective is *maximum weighted likelihood estimation* [26], which consists in maximizing the sum of the log-density of each sample item, suitably weighted by a coefficient $\omega \in \mathbb{R}_0^+$: the higher the coefficient and more influential will be the corresponding sample point in the overall optimization. Here we propose two variants of the above vanilla regression models, by introducing two weighting schemes $\boldsymbol{\omega}^a$ and $\boldsymbol{\omega}^b$, as follows. Having denoted with

$n_+$ and $n_-$ respectively the number of positive and negative instances:

$$\omega_i^a = \begin{cases} 1/n_+ & \text{if gene } i \in S_+ \\ 1/n_- & \text{otherwise} \end{cases} \qquad \omega_i^b = \begin{cases} \Delta_i^+ / \sum_{j \in S_+} \Delta_j^+ & \text{if gene } i \in S_+ \\ 1/n_- & \text{otherwise} \end{cases} \tag{2}$$

Intuitively, both schemes try to compensate the class imbalance by giving higher weights to infrequent instances. Scheme 'b' breaks the flatness of positive weights by assigning higher influence to positive nodes when they are functionally more similar to nodes belonging to the same class. In other words, the higher is the positive neighborhood of a positive node and the higher will be its influence in the overall maximization process.

Summing up all possible combinations of GLMs and weight schemas, we obtained a total of 12 models, which we refer to with the schema "[W]GP-*mod* [*ws*]", where WGP stands for Weighted Gene Prioritization, *mod* is one of the four GLM acronyms used in Table 2, the weights schema *ws* ∈ {'a','b'}, and square brackets are used to denote optional arguments.

## 5 Results and Discussion

In order to have a fair comparison, the experimental validation of the proposed models follows the setting adopted in [9]. We compared our method with the state-of-the-art techniques briefly described in Table 3, and estimated the generalization performances by averaging the performances observed through the classical $k$-fold cross-validation (CV), with $k = 5$. Performances have been assessed using both the *Area Under the ROC Curve* (AUC) and the *Precision* at different *Recall* levels (PXR). Concerning the experimental campaign, as performed in [9], firstly we run our methods on the network Net6 (see Sect. 3.1). Table 4 shows the corresponding average AUC sorted in decreasing order. Apart from GP-PR and GP-CLogLR, all our methods outperform the top-performing benchmark algorithm ($S_{AV}$ $t = 5$). This witnesses the high informativeness of the two projected features defined in Eq. (1) and the effectiveness of the GLM to cope with the label imbalance at each node neighborhood. Moreover, to appreciate the benefit of the cost-sensitive models w.r.t. the corresponding vanilla versions, we performed the one-side Wilcoxon Signed Rank test between all couples of methods within the same family to assess whether their population mean ranks differ. As a results, we observed a meaningful increase in performance of the 'b' scheme over the 'a' variant – confirming our initial assumption that positives, carrying more information than negatives, should be taken into account when learning the predictive model – and singularly of both cost-sensitive models w.r.t. their vanilla version ($p$-value $< 0.001$). This regularity breaks down in both linear and Poisson regressions, where scheme 'a' outperforms variant 'b' ($p$-value $= 0.025$). We conjecture that such results are due to the convergence of GLM fitting procedures toward spurious optima in rare instances which, in turn, may be caused by the peaked landscape of weights distribution in 'b' scheme. Finally, due to the fast convergence of regression performed in the 2-dimensional projected space,

**Table 3.** Competitor benchmark methods.

| |
|---|
| *Kernelized score functions.* This kernel based ranking method adopts a suitable kernel matrix so as to extend the similarity between two nodes $i$ and $j$ also to non neighboring nodes [9]. The score of each gene $i$ for a given MeSH disease $M$ is defined according to a suitable metric $d(i, V_M)$, which is specified in terms of a distance $d_{\mathcal{H}}$ between the images in a suitably chosen Hilbert space $\mathcal{H}$ of $i$ and the subset of genes $V_M \subset V$ associated with $M$. By varying the definition of $d(i, V_M)$, authors obtained different scoring methods:<br>– $\mathrm{S_{NN}}$, when $d(i, V_M)$ is the minimum distance (in $\mathcal{H}$) between $i$ and $V_M$;<br>– $\mathrm{S_{kNN}}$, when $d(i, V_M)$ is defined by considering the closest $k$ neighbors in $V_M$;<br>– $\mathrm{S_{AV}}$, when $d(i, V_M)$ is the average distance between $i$ and $V_M$.<br>As kernel matrix, the $t$-step ($t = 1,\ 2,\ \ldots$) random walk kernel $\boldsymbol{K}^t$ is adopted, where $\boldsymbol{K} = \gamma \boldsymbol{I} + \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{W} \boldsymbol{D}^{-\frac{1}{2}}$, $\boldsymbol{I}$ is the $n \times n$ identity matrix, and $\gamma > 0$. |
| *Random walks.* The classical $t$-step random walk (RW) algorithm [27] assigns to a node $i \in V$ a score corresponding to the probability that a $t$-step random walk in $G$, starting from positive nodes ends at node $i$. The transition matrix $\boldsymbol{T}$ adopted by the random walker is obtained from $\boldsymbol{W}$ by row normalization, that is $\boldsymbol{T} = \boldsymbol{D}^{-1} \boldsymbol{W}$. |
| *Random walks with restart.* The rationale behind the random walk with restart (RWR) algorithm is that after many steps the walker may forget the prior information coded in the initial probability vector (0 for nodes in $V \setminus V_M$ and $1/\lvert V_M \rvert$ for nodes in $V_M$, for MeSH term $M$). Thus, the algorithm allows the walker to move another random walk step with probability $1 - \theta$, or to restart from its initial condition with probability $\theta$. |
| *Guilt-by-association methods.* Algorithms relying upon the GBA rule make predictions based on the interacting genes, which are assumed to share more likely similar functions [28]. Usually, the discriminant score for a gene $i$ w.r.t. a given MeSH disease $M$ is obtained as sum of the weights connecting $i$ to neighboring genes associated with the disease $M$, or as the maximum of these weights. The benchmark results adopt the latter version. |

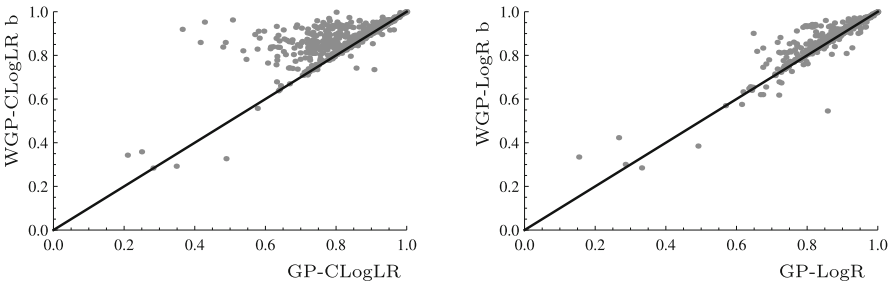
our method is also scalable, taking around 5 seconds to perform the entire 5-fold CV procedure for a single MeSH disease on a Intel i7-860 CPU 2.80 GHz machine with 16 GB of RAM.

To better investigate the improvements achieved by cost-sensitive methods, we show in Fig. 1 the paired AUC obtained by vanilla regression and the corresponding cost-sensitive 'b' version for cloglog and logit link functions (similar trends were obtained for all other paired comparisons – results not shown). It is immediate to observe a large majority of bullets lying above the bisector, showing that cost-sensitive variants achieve higher AUC values for most of the considered MeSH diseases. Indeed, in the first two columns of Table 5 we report the proportion of MeSH terms where cost-sensitive methods outperform the corresponding vanilla ones. Such proportion ranges from 70.1 % to 85.9 %.

Similar AUC results are obtained when running the proposed methods over the network Net9, as reported in Table 6. Results obtained by GBA, RW and RWR methods are not reported in the referenced papers due to their low performances. All our methods (except for GP-CLogLR) perform better than the top-performing benchmark method ($\mathrm{S_{AV}}$ $t = 5$). Note how the best method (GP-CLogLR b) makes more noticeable the gain due to the adopted cost-sensitive approach, ranking the correspondent vanilla version at thirteenth place. The Wilcoxon Signed Rank test confirms the results observed for the network Net6,

**Table 4.** Average AUC across MeSH terms on the network Net6.

| Method | AUC | Method | AUC | Method | AUC |
|---|---|---|---|---|---|
| WGP-CLogLR b | 0.8777 | RWR $\theta = 0.6$ | 0.8565 | $S_{kNN}$ $t = 1, k = 19$ | 0.8138 |
| WGP-LogR b | 0.8767 | GP-PR | 0.8563 | RW $t = 3$ | 0.7937 |
| WGP-CLogLR a | 0.8762 | $S_{AV}$ $t = 2$ | 0.8562 | RW $t = 5$ | 0.7773 |
| WGP-LR b | 0.8757 | $S_{AV}$ $t = 10$ | 0.8548 | RW $t = 10$ | 0.7720 |
| WGP-LR a | 0.8748 | $S_{AV}$ $t = 1$ | 0.8538 | $S_{kNN}$ $t = 10, k = 3$ | 0.7636 |
| WGP-LogR a | 0.8737 | RWR $\theta = 0.6$ | 0.8533 | $S_{kNN}$ $t = 5, k = 3$ | 0.7405 |
| GP-LR | 0.8705 | $S_{kNN}$ $t = 10, k = 19$ | 0.8374 | $S_{kNN}$ $t = 3, k = 3$ | 0.7332 |
| WGP-PR a | 0.8680 | GP-CLogLR | 0.8365 | $S_{kNN}$ $t = 2, k = 3$ | 0.7304 |
| WGP-PR b | 0.8665 | GBA | 0.8313 | $S_{kNN}$ $t = 1, k = 3$ | 0.7280 |
| GP-LogR | 0.8648 | $S_{kNN}$ $t = 5, k = 19$ | 0.8251 | $S_{NN}$ $t = 10$ | 0.7251 |
| $S_{AV}$ $t = 5$ | 0.8596 | $S_{kNN}$ $t = 3, k = 19$ | 0.8199 | $S_{NN}$ $t = 5$ | 0.7020 |
| $S_{AV}$ $t = 3$ | 0.8580 | RW $t = 2$ | 0.8186 | $S_{NN}$ $t = 3$ | 0.6968 |
| RW $t = 1$ | 0.8566 | $S_{kNN}$ $t = 2, k = 19$ | 0.8170 | $S_{NN}$ $t = 2$ | 0.6950 |
| | | | | $S_{NN}$ $t = 1$ | 0.6934 |




(a) (b)

**Fig. 1.** Paired AUC comparison between cost-sensitive 'b' schema and the corresponding vanilla version for: (a) cloglog and (b) logit link functions.

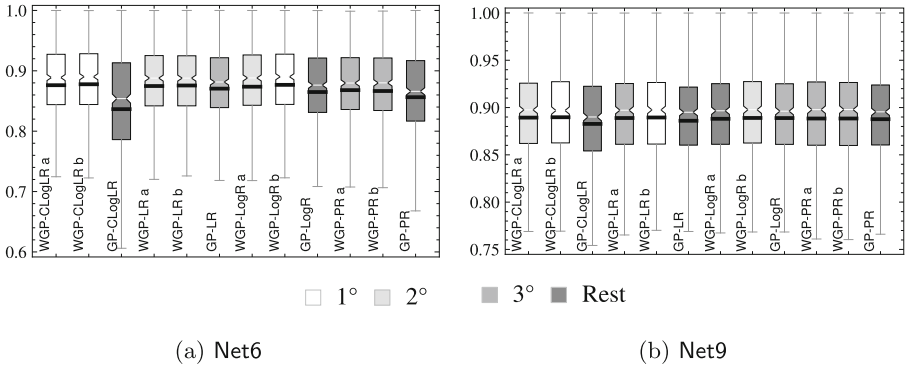
**Table 5.** Proportion of wins (in terms of AUC) of cost-sensitive vs. cost-insensitive methods, observed over all the considered MeSH terms.

| | Network Net6 | | Network Net9 | |
|---|---|---|---|---|
| | WGP a | WGP b | WGP a | WGP b |
| GP-LR | 0.701 | 0.709 | 0.688 | 0.756 |
| GP-LogR | 0.743 | 0.804 | 0.441 | 0.579 |
| GP-CLogLR | 0.833 | 0.859 | 0.610 | 0.638 |
| GP-PR | 0.768 | 0.732 | 0.513 | 0.535 |

**Table 6.** Average AUC across MeSH terms on the network Net9.

| Method | AUC | Method | AUC | Method | AUC |
|---|---|---|---|---|---|
| WGP-CLogLR b | 0.8897 | GP-PR | 0.8877 | $S_{kNN}$ $t = 5, k = 19$ | 0.8500 |
| WGP-LR b | 0.8895 | GP-LR | 0.8860 | $S_{kNN}$ $t = 3, k = 19$ | 0.8413 |
| WGP-CLogLR a | 0.8894 | $S_{AV}$ $t = 5$ | 0.8831 | $S_{kNN}$ $t = 2, k = 19$ | 0.8368 |
| WGP-LogR b | 0.8890 | GP-CLogLR | 0.8827 | $S_{kNN}$ $t = 1, k = 19$ | 0.8322 |
| WGP-LR a | 0.8889 | $S_{AV}$ $t = 3$ | 0.8811 | $S_{NN}$ $t = 10$ | 0.7437 |
| GP-LogR | 0.8889 | $S_{AV}$ $t = 2$ | 0.8792 | $S_{NN}$ $t = 5$ | 0.7106 |
| WGP-PR b | 0.8884 | $S_{AV}$ $t = 1$ | 0.8765 | $S_{NN}$ $t = 3$ | 0.7014 |
| WGP-PR a | 0.8884 | $S_{AV}$ $t = 10$ | 0.8761 | $S_{NN}$ $t = 2$ | 0.698 |
| WGP-LogR a | 0.8881 | $S_{kNN}$ $t = 10, k = 19$ | 0.8665 | $S_{NN}$ $t = 1$ | 0.695 |

with some exceptions. Firstly, we observe no meaningful differences between both the two cost-sensitive variants of the Poisson model, and 'a' schema with its naive version ($p$-value $> 0.05$). Moreover, the only model privileging the vanilla variant w.r.t. its 'a' schema counterpart is the logistic one ($p$-value $< 0.001$). The exceptional nature of such an event is confirmed by the entries reported in the last two columns of Table 5: despite the less pronounced proportion of wins of cost-sensitive methods over their cost-insensitive variants than those observed for network Net6, six out of eight entries still shows a remarkable disproportion in favor of cost-sensitive schemas.




(a) Net6 (b) Net9

**Fig. 2.** Performance distribution of the proposed methods across MeSH terms for: (a) network Net6, and (b) network Net9. Box colors depict the performance ranking of each method, as explained by the legend reported below the graphs. In such setting, boxes sharing the same colors represent indistinguishable methods.

To better analyse the AUC distributions over MeSH diseases, in Fig. 2 we report the box-and-whiskers plot of all proposed methods. Boxes are colored so

as to reflect the ranking of the methods, obtained by performing all pairwise comparisons under the one-side Wilcoxon Signed Rank test. Models sharing the same color represent maximal sets of indistinguishable methods under the above test with 0.05 significance level. The darker the color, the worst is the ranking, as shown in the legend under the picture. In particular, all methods ranking fourth downward are joined together in the same class, for the sake of visualization. Apart from the already discussed over-performance of cost-sensitive methods, we appreciate both a smaller variance and a reduced presence of outliers (not shown in the pictures). It is worth noting the marked skewness toward lower AUC values in all experiments, as confirmed by the fact that the means of AUC distributions (black markers in the pictures) are always lower than their medians (depicted with notches). Evaluating performances through means, as done in Tables 4 and 6, strongly penalizes all methods, being mean values strongly affected by the presence of outliers having low AUC values. To guarantee a fair comparison with benchmark results, we still make use of such estimator, noting that median values give a more informative and less biased view of the overall performances.

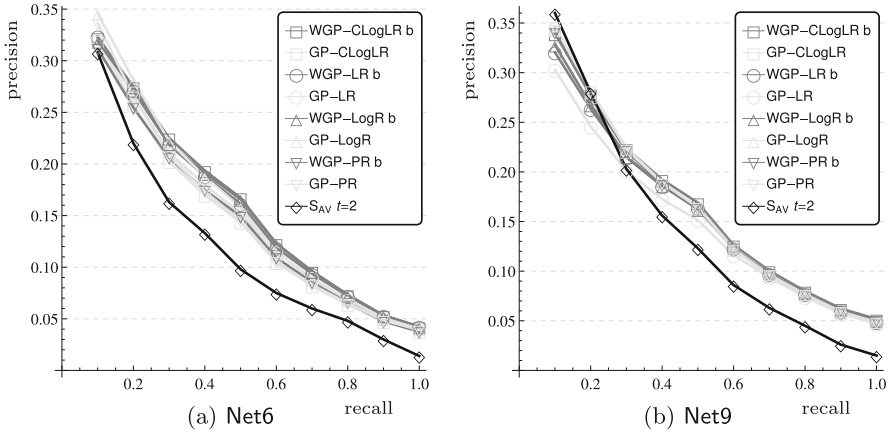



**Fig. 3.** PXR levels achieved on the network: (a) Net6, and (b) Net9.

We conclude this analysis by showing in Fig. 3 the PXR results for recall levels ranging from 0.1 to 1, with steps of 0.1. Undoubtedly, the performances of the proposed methods are very close each others; for this reason, for the sake of readability, we report just the results for vanilla and cost-sensitive 'b' scheme methods, since 'a' scheme achieves almost indistinguishable results. To better appreciate the advantage of working with cost-sensitive methods, we use a light gray level for all vanilla methods, and a dark gray one for their cost-sensitive variants. Apart from the slight but always remarkable behavior of the latter, we observe a noticeable gain w.r.t. $S_{AV}$ $t = 2$, the only method of which authors published the PXR performances, with the exception of precision value at a recall level of 0.1 in picture (b), where light and dark gray lines are almost overlapped,

apart from GP-LR method which performs slightly worse. Note that in Fig. 3(a), vanilla methods tend to be more accurate for lower levels of recall. Nevertheless, for all the remaining recall values, in particular in the range $[0.3, 1]$, cost-sensitive methods always outperform cost-insensitive ones.

## 6 Conclusions

In this work we propose a novel approach for gene-disease prioritization which is specifically designed to deal with imbalanced data, such as those characterizing databases of seed genes for known human diseases. We have shown that imbalance-aware methods can noticeably improve the performance in detecting gene-disease associations, evaluating the effectiveness of the proposed approach on a larged sized benchmark for gene prioritization problem. Future works will be devoted to exploit the hierarchical contribution coming from ontologically related gene classes.

## References

1. Lehne, B., Lewis, C.M., Schlitt, T.: From SNPs to genes: disease association at the gene level. PLoS ONE **6**(6), e20133 (2011)
2. Manolio, T.A.: Genomewide association studies and assessment of the risk of disease. N. Engl. J. Med. **363**(2), 166–176 (2010)
3. Brnigen, D., et al.: An unbiased evaluation of gene prioritization tools. Bioinformatics **28**(23), 3081–3088 (2012)
4. Ashburner, M., et al.: Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat. Genet. **25**(1), 25–29 (2000)
5. Navlakha, S., Kingsford, C.: The power of protein interaction networks for associating genes with diseases. Bioinformatics **26**(8), 1057–1063 (2010)
6. Vanunu, O., Sharan, R.: A propagation-based algorithm for inferring gene-disease associations. In: Proceedings of the German Conference on Bioinformatics, GCB, September 9–12, Dresden, Germany (2008)
7. Kohler, S., et al.: Walking the interactome for prioritization of candidate disease genes. Am. J. Hum. Genet. **82**(4), 949–958 (2008)
8. Antanaviciute, A., et al.: Ova: integrating molecular and physical phenotype data from multiple biomedical domain ontologies with variant filtering for enhanced variant prioritization. Bioinformatics **31**(23), 3822–3829 (2015)
9. Valentini, G., et al.: An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. Artif. Intell. Med. **61**(2), 63–78 (2014)
10. Frasca, M., et al.: UNIPred: unbalance-aware network integration and prediction of protein functions. J. Comput. Biol. **22**(12), 1057–1074 (2015)
11. Amberger, J., Bocchini, C., Hamosh, A.: A new face and new challenges for online mendelian inheritance in man (OMIM). Hum. Mutat. **32**(5), 564–567 (2011)
12. Elkan, C.: The foundations of cost-sensitive learning. In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, pp. 973–978 (2001)
13. Frasca, M., et al.: A neural network algorithm for semi-supervised node label learning from unbalanced data. Neural Netw. **43**, 84–98 (2013)

14. Frasca, M.: Automated gene function prediction through gene multifunctionality in biological networks. Neurocomputing **162**, 48–56 (2015)
15. Bertoni, A., Frasca, M., Valentini, G.: COSNet: a cost sensitive neural network for semi-supervised learning in graphs. In: Hofmann, T., Malerba, D., Vazirgiannis, M., Gunopulos, D. (eds.) ECML PKDD 2011, Part I. LNCS, vol. 6911, pp. 219–234. Springer, Heidelberg (2011)
16. Frasca, M., Pavesi, G.: A neural network based algorithm for gene expression prediction from chromatin structure. In: IEEE IJCNN, pp. 1–8 (2013). doi:10.1109/IJCNN.2013.6706954
17. Davis, A.P., et al.: Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic Acids Res. **37**(Database issue), D786–D792 (2009)
18. Wu, G., Feng, X., Stein, L.: A human functional protein interaction network and its application to cancer data analysis. Genome Biol. **11**(5), R53+ (2010)
19. Lee, I., et al.: Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. **21**(7), 1109–1121 (2011)
20. Segal, E., et al.: A module map showing conditional activity of expression modules in cancer. Nat. Genet. **36**(3), 1090–1098 (2004)
21. Chatr-aryamontri, A., et al.: The biogrid interaction database: 2013 update. Nucleic Acids Res. **41**(Database–Issue), 816–823 (2013)
22. Hellevik, O.: Linear versus logistic regression when the dependent variable is a dichotomy. Qual. Quant. **43**(1), 59–74 (2009)
23. Van Del Paal, B.: A comparison of different methods for modelling rare events data. Master thesis in statistical data analysis, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium (2013–2014)
24. Derby, N.: An introduction to the analysis of rare events. In: SA16 Proceedings of the 2011 Midwest SAS Users Group Conference, Kansas City, KS (2011)
25. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
26. Dmochowski, J.P., Sajda, P., Parra, L.C.: Maximum likelihood in cost-sensitive learning: model specification, approximations, and upper bounds. J. Mach. Learn. Res. **11**, 3313–3332 (2010)
27. Lovász, L.: Random walks on graphs: a survey. In: Miklós, D., Sós, V.T., Szőnyi, T. (eds.) Combinatorics, Paul Erdős is Eighty, vol. 2, pp. 353–398. János Bolyai Mathematical Society, Budapest (1996)
28. Schwikowski, B., Uetz, P., Fields, S.: A network of protein-protein interactions in yeast. Nat. Biotechnol. **18**(12), 1257–1261 (2000)