<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Ensemble positive unlabeled learning for disease gene identification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Yang, Peng; Li, Xiaoli; Chua, Hon-Nian; Kwoh, Chee-Keong; Ng, See-Kiong</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>Yang, P., Li, X., Chua, H.-N., Kwoh, C.-K., &amp; Ng, S.-K. (2014). Ensemble Positive Unlabeled Learning for Disease Gene Identification. PLoS ONE, 9(5), e97079-</td>
</tr>
<tr>
<td><strong>Date</strong></td>
<td>2014</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10220/19767">http://hdl.handle.net/10220/19767</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>© 2014 Yang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.</td>
</tr>
</tbody>
</table>
Ensemble Positive Unlabeled Learning for Disease Gene Identification

Peng Yang1, Xiaoli Li1, Hon-Nian Chua1, Chee-Keong Kwoh2, See-Kiong Ng1

1 Data Analytics Department, Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore
2 Bioinformatics Research Centre, School of Computer Engineering, Nanyang Technological University, Singapore, Singapore

Introduction

While high-throughput genomic studies have led to the discovery of hundreds and thousands of candidate disease genes, the identification of genes involved in specific human diseases has remained a fundamental challenge, requiring time-consuming and expensive experimentation. Computational approaches that can reliably predict novel disease genes from the vast number of unknown genes will provide a useful alternative to speed up the long and arduous searches for the genetic causes of various human disorders.

Given that an increasing number of genes have been experimentally confirmed over the years as causative genes to various human diseases, it will be useful to develop machine learning methods to identify novel disease genes from the confirmed disease genes as positive training examples, based on the observation that genes associated with similar disease phenotypes are likely to share similar biological characteristics. For example, proteins involved in hereditary diseases tend to be long, with more homologs with distant species, but fewer paralogs within human genome [1]. They are also likely to attach together to form functional modules such as protein complexes [2]. In fact, various studies have shown that genes associated with similar disorders tend to demonstrate similar gene expression profiling [3], high functional similarities [4] [5] and physical interactions between their gene products [6] [7].

In addition, with disease phenotype similarity data, genes associated with same/similar disease phenotypes are likely to share similar biological functions. Given a phenotype \(\phi_h\), we can infer its potential disease genes from those disease genes associated with phenotypes \(\phi_h/\phi_h\) and \(\phi_h\) are very similar) [8].

A number of methods above have thus been proposed to prioritize candidate genes based on different kinds of biological data, such as gene sequence data, gene expression profile, evolutionary features, functional annotation data and PPI dataset. Adie et al. [9] employed a decision tree algorithm based on a variety of genomic sequence and evolutionary features, such as coding sequence length and evolutionary conservation, presence, and closeness of paralogs in the human genome. Topological information on PPI network has also been demonstrated to be useful for disease gene prediction. Smalter et al. [10] applied support vector machines (SVM) classifier using PPI topological features in addition to sequence-derived and evolutionary features, while Radivojac et al. [11] built three individual SVM classifiers using three types of features—PPI network, protein sequence and protein functional information—and then built a final classifier to...
combine the predictions from three individual classifiers for candidate gene prediction.

The research work mentioned above employed classical machine learning methods to build a binary classifier where the confirmed disease genes are used as positive training set $P$ and unknown genes as negative training set $N$. However, these machine learning techniques hardly perform as well as they could because the negative set $N$ that they used contained unconfirmed disease genes (false negatives). In light of aforementioned limitation, recently positive unlabeled learning (PU learning) methods have been proposed for the task by building a classification model in which unknown genes are appropriately treated as an unlabeled set $U$ (instead of a negative set $N$). For example, Mordelet et al. proposed a bagging method ProDiGe for disease gene prediction. It iteratively choose random subsets (RS) from $U$ and then trained multiple classifiers using bias SVM to discriminate $P$ from each subset RS. The multiple classifiers were subsequently aggregated to generate the final classifier [12]. Given that the RS’s were likely to contain less noise (unknown disease genes) than the original set $U$, it was able to perform better than classical binary classification models that inappropriately used $N$ as negative training data. More recently, Yang et al. designed a novel multi-level PU learning algorithm PUDI to build a classifier with better performance for disease gene identification where the unlabeled set $U$ was partitioned into multiple positive and negative sets with confidence scores for building the classifier [13].

The prior works have clearly shown that integration of various biological data sources is not only desirable but also essential for robust disease gene prediction, since using only a single source of data for prediction is susceptible to incompleteness and noise in the genomic data. It is also advantageous to employ an ensemble approach for prediction, since using a single machine learning predictor is similar in risk of potential bias caused by inherent limitations of individual prediction models. In this paper, we propose an effective PU learning framework to integrate multiple biological data sources and an ensemble of powerful machine learning classifiers for disease gene identification. In our proposed framework, we first extract multiple positive and negative samples from unlabeled set $U$ through performing random walk with restart on different biological networks. We use three biological networks for this paper: protein interaction network, gene expression similarity network, and GO similarity network [16] [17] [18] are constructed. Then we will present the schema of our proposed EPU algorithm.

Experimental data and gene network modeling

In this paper, we have exploited the following biological data human protein interaction data, gene expression data, gene ontology, and phenotype-gene association data.

Human protein interaction data (PPI) is downloaded from the Human Protein Reference Database (HPRD) [19] and Online Predicted Human Interaction Database (OPHD) [20]. The combined PPI dataset contains 143,939 PPIs among a total of 13,035 human proteins. We build a protein interaction network $G_{PPI} = \{V_{PPI}, E_{PPI}\}$ where $V_{PPI}$ represents the set of vertices (proteins) and $E_{PPI}$ denotes all edges (detected pairwise interactions between proteins). $G_{PPI}$ can be represented as its matrix format, i.e. $W_{PPI} = [w_{ij}]$, where $w_{ij} = 1$ if the corresponding protein pair $(v_i, v_j) \in E_{PPI}$, 0 otherwise.

Gene expression data is obtained from RNASeq data which is made publicly available in the EBI ArrayExpress, by the Illumina Human BodyMap 2.0 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30611). The dataset comprises Fastq reads from the paired-end sequencing of cells from 16 human tissue types, including colon, heart, kidney, white blood cells and so on, using the Illumina HiSeq next generation sequencing platform. This dataset represents the expression values of 17,652 human genes on 16 human tissue types. Suppose gene $g_i$ and $g_j$ are represented as their gene expression profile vectors $(x_{i1}, x_{i2}, ..., x_{in})$ and $(x_{j1}, x_{j2}, ..., x_{jn})$ respectively where $x_{ik}$ $(k = 1, 2, ..., n)$ denotes the expression value of gene $i$ in the $k$-th tissue. Pearson correlation coefficient is employed to measure the similarity between $g_i$ and $g_j$:

$$sim_{GE}(g_i, g_j) = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^{n}(x_{jk} - \bar{x}_j)^2}}$$ (1)

where $\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n}x_{ik}$, $\bar{x}_j = \frac{1}{n}\sum_{k=1}^{n}x_{jk}$.

We build a gene expression similarity network $G_{GE} = \{V_{GE}, E_{GE}\}$, where $V_{GE}$ represents a set of genes occurring in the gene expression data and $E_{GE}$ represents a set of edges between the genes in $V_{GE}$. For each gene $g_i$, we ranked all genes $g_j$ in $V_{GE}$ where $i \neq j$ according to decreasing order of $sim_{GE}(g_i, g_j)$, and add an edge $(g_i, g_j)$ to $E_{GE}$ if $g_j$ is in the top 5 of the list. This helps to filter low similarity pairs and potential noise in gene expression data. Then we transform the gene expression network $G_{GE}$ to its matrix format where the edges of two rows are reformatted to their gene expression similarity in equation (1).

Gene Ontology (GO, http://www.geneontology.org/) is a set of controlled vocabulary used to annotate genes and gene products [21]. Gene Ontology provides three sub-ontologies, namely, biological process (BP), molecular function (MF) and cellular component (CC) [21]. For each gene, we build a feature vector using its biological annotations from these three sub-ontologies, i.e. $\{MF_1, ..., MF_{|SMF|}, BP_1, ..., BP_{|BP|}, CC_1, ..., CC_{|CC|}\}$. For example, a gene $g_i$ is represented as gene vector $g_i = (m_{i1}, ..., m_{i|SMF|}, b_{i1}, ..., b_{i|BP|}, c_{i1}, ..., c_{i|CC|})$, where $m_{ij}$ (similar for $b_{ip}$, $c_{iq}$) is GO term similarity between $g_i$ and the feature $MF_j$. Since the GO terms of BP, MF and CC are organized into DAG structure, we use the computational method in [22] to measure the similarity of two GO terms. And $|SMF|$ is number of selected MF term features.
We choose the GO features that could help distinguish disease genes from non-disease genes with strategy in [13] and top 1000 scored features were selected for each of three feature groups, i.e. BP, MF and CC, respectively. We then build GO similarity network $G_{GO}=(V_{GO}, E_{GO})$, where $V_{GO}$ is the gene set annotated in GO dataset and $E_{GO}$ is a set of edges between the genes in $V_{GO}$. Similarly to the gene expression similarity network, we keep those top 5 edges which have highest similarities to each gene and other edges. $G_{GO}$ can be represented as its matrix format, i.e. $W_{GO}=[w_{ij}]$. Given a gene $g_k$, if $g_k$ is one of top 5 lists of $g_j$, $w_{kj}$ is normalized as:
\[
w_{kj} = \frac{\text{Dis}(g_k,g_j) - \min_{g_j \in V_{GO}} \text{Dis}(g_k,g_j)}{\max_{g_j \in V_{GO}} \text{Dis}(g_k,g_j) - \min_{g_j \in V_{GO}} \text{Dis}(g_k,g_j)}
\]
otherwise, $w_{kj} = 0$, where $\text{Dis}(g_k,g_j)$ denotes Euclidean distance between $g_k$ and $g_j$ and $0 \leq w_{kj} \leq 1$.

**Phenotype-gene association data.** 4260 phenotype-gene association data, spanning 2659 known disease genes and 3200 disease phenotypes, are obtained from the latest version of OMIM (http://omim.org/) [23]. Goh et al. [6] have categorized the 3200 disease phenotypes in OMIM database into 22 disease groups, i.e. Cancer, Metabolic, Neurological, Endocrine, etc., based on the physiological system affected. For example, the Endocrine disease group comprises 62 OMIM phenotypes, including OMIM 241830 (Bamforth-Lazarus syndrome) and OMIM 304800 (Diabetes insipidus, nephrogenic) etc.

**Phenotype similarity network.** Disease phenotype similarity network [24], is defined as $G_{PH}=(V_{PH}, E_{PH})$, where $V_{PH}$ denotes the set of disease phenotypes and $E_{PH}$ denotes relevant phenotype pairs. Disease phenotypes in $V_{PH}$ are represented as feature vectors in which feature terms are Medical Subject Headings (MeSH) controlled vocabulary, and phenotype similarities in $E_{PH}$ are evaluated underline concept relevance and frequency of MeSH terms appearing in text description of OMIM documents. According to Vanunu et al. [8], phenotype pairs with high similarities are regarded as informative and reliable. Therefore, we apply logistic function to filter out low phenotypic similarities in $E_{PH}$, following [2] [8].

**The proposed technique EPU**

The schema of our EPU algorithm is presented in Figure 1. EPU first selects candidate positives from positive genes and reliable negatives from unlabeled genes. It then builds three gene similarity networks using PPI data, gene expression data and Gene Ontology data and applies random walk on the three networks to propagate weights to unlabeled genes that reflect likelihoods of belonging to positive/negative class. We then exploit the weighted genes to build three diverse classification models to predict “soft” labels for test genes. Finally, an ensemble learning algorithm combines the prediction results from the classifiers to make a final prediction for the classification of the unknown test gene.

Suppose all disease genes from OMIM are stored into a disease gene set $DIS$. All the other genes that are not a member of $DIS$ will be treated as unknown/unlabeled genes and be stored into a set $UG$ (contains 16,570 genes) [25]. Each gene in $DIS$ and $UG$ is represented as a feature vector, namely, $\mathbf{f} = \{f_1, ..., f_m\}$ where $m$ is the total number of features from GO terms, protein domains and PPI topological features, following our previous work [13].

In the next section, we describe how to predict novel disease genes given the confirmed disease genes for a particular disease or disorder. The confirmed disease genes for the given disorder group are treated as positive set $P$ ($P \subset DIS$) while randomly selected unknown genes from $UG$ are treated as unlabeled set $U$ ($U \subset UG$, $|U| = |P|$), following the settings in [9] [10] [13].

**Weighting unlabeled genes by integrating multiple biological evidences**

Given a particular disease class and its known associated disease genes, we first build the training data sets for machine learning by prioritizing the candidate positives and reliable negatives based on their similarity to the query disease class. We build three gene similarity networks using PPI, gene expression and GO as described above, and perform a random walk with restart algorithm on these three gene similarity networks to estimate the likelihood of the unlabeled genes belonging to disease class or non-disease class. The details are as follows.

**Extracting candidate positives and reliable negatives.** As a typical positive set $P$ is relatively small, we want to find a set of candidate positive genes $CP$ to complement $P$. Given that recent studies have shown that similar phenotypes are often caused by functionally related disease genes [4] [6], we could populate the set of candidate positive genes $CP$ with genes associated to similar/relevant phenotypes, based on the principle of guilt-by-association. In other words, given a query disease group/class, we can use its associated phenotypes to uncover similar disease phenotypes, as shown in Figure S1.

Having identified the candidate positive genes $CP$, let us now describe how to extract reliable negative gene set $RN$. We consider reliable negatives as those unlabeled genes that are very different from positive set $P$. To identify such genes, we first build a “positive representative vector” $\tilde{p}$ by summing up gene vectors in $P$ and normalizing it. Then, we compute the average Euclidean distance [26] of every unlabeled gene in $U$ from $pr$. To extract the reliable negative genes for $RN$, we regard an unlabeled gene $g_k$ as a member of $RN$ if its distance from $pr$ is longer than the average distance (of all the genes in $U$) from $pr$, formalized as follows:

$$RN = \{g_k | \text{dis}(pr,g_k) > D\}$$

where $\text{dis}(pr,g_k)$ is the Euclidean distance between $pr$ and positive representative vector $pr$. We compute an average distance $D$ of all the unlabeled gene in $U$ from $pr$ as $D = \frac{1}{|U|} \sum_{g_j \in U} \text{dis}(pr,g_j)$.

**Ensemble weighted unlabeled genes via performing label propagation on multiple networks.** We now have the given positive set $P$, a candidate positive set $CP$, a reliable negative set $RN$ and a remaining unlabeled set $U = U - RN$ for machine learning. To build a robust classification model, we will extract those genes with reliable labels that are near the decision boundary between the positive and negative classes. We adapt the Random Walk with Restart algorithm [27] to perform flow propagation which spreads the label information from $P$, $CP$ and $RN$ to the unlabeled genes in $U$ on the biological networks that we have constructed, namely the PPI network $G_{PP}$, the GO similarity network $G_{GO}$ and the gene expression similarity network $G_{PH}$ as described earlier.

Formally, let $R_0$ be an initialization vector where primitive scores are assigned to all genes in three networks to indicate the genes’ potential classification labels. Let $p_0$, $P_0$ and $n_0$ denote the initial values for genes in $P$, $CP$ and $RN$ respectively, as follows. The genes $g \in P$ are all given a score $p_0(g) = +1$, indicating their disease gene status. Each candidate positive gene $g \in CP$ is assigned a score that computes its maximal phenotypic similarity to the known disease genes in $P' = \max_{p \in CP} \text{PH}(g, p)$, where $\text{PH}(g)$ denotes disease phenotypes caused by gene $g$, and $PH(P)$...
denotes disease phenotypes caused by disease set $P$. For genes in reliable negative set $RN$, to balance total amount of flows between positive genes and negative genes, the initial score for negative gene $g_i$ is assigned with

$$n_0(g_i) = \frac{-\sum_{g_j \in P} p_0(g_i) + \sum_{g_j \in CP} p'_0(g_i)}{|RN|}$$

where $\sum_{g_j \in P} p_0(g_i) + \sum_{g_j \in CP} p'_0(g_i)$ is a total amount of positive gene set. The remaining unlabeled genes in $U'$ are assigned an initial score of 0.

For each of the three biological networks $G_{PPI}$, $G_{GE}$ and $G_{GO}$, prior influence from the seed nodes in $P$, $CP$ and $RN$ are propagated to their direct neighbors, and then continue to spread to other adjacent nodes iteratively across the network. Given $R_0$, the initial score vector (step 0), $R_t$, the score vector at step $t$, can be calculated as follows:

$$R_t = (1 - z)WR_{t-1} + zR_0, (t \geq 2)$$

where $R_t = R_0$ and $W = D^{-1/2}W$ is a normalized format of matrix $W$. $W \in \{W_{PPI}, W_{GO}, W_{GE}\}$. Here $D$ is the diagonal matrix and $D_{ii} = \sum_k W_{ik}$. $z$ represents the percentage of flow back to original seed nodes in $P$, $CP$ and $RN$ during each iteration. The default value of 0.7 is used for $z$, following [16].

Eventually, the information flow will converge to a steady state [27]. In our case, the Random Walk with Restart algorithm will stop its iterative process when the difference between two steps $R_t$ and $R_{t-1}$ is less than $10^{-6}$ [16], measured by $L1$ norm. The scores for unlabeled genes from the three gene networks are combined into one integrated score:

$$Int_{score}(g) = \frac{1}{3} (R_t(g, W_{PPI}) + R_t(g, W_{GO}) + R_t(g, W_{GE}))$$
where $R_{g,W_PPI}$, $R_{g,W_GO}$ and $R_{g,W_GE}$ are the scores for gene $g$ in the PPI, GO similarity and gene expression similarity networks respectively.

**Ensemble positive unlabeled learning EPU**

Next, we describe how to build three separate PU learning classification models Support Vector Machine SVM, K-Nearest Neighbor, and Naive Bayes classifier—to classify genes into two classes $C=\{+,-\}$, where ‘+’ denotes positive/disease class and ‘−’ presents negative/non-disease class.

**PU learning model 1: Weighted K-Nearest Neighbor (WKNN).** KNN is an instance based learning method, which classifies an unknown test gene based on the class labels of its top $K$ nearest training example genes, i.e. based on the majority class vote of its nearest $K$ neighbors. The distance between the test gene and other training examples can be computed using common distance metrics such as Euclidean distance. Given a test gene $g_t$ and its $K$ nearest neighbor set $D_k$ we divide $D_k$ into positive and negative training subsets, namely $D_p = \{g | \text{Int}_\text{score}(g) \geq 0, \ g \in D_k\}$ and $D_0 = \{g | \text{Int}_\text{score}(g) < 0, \ g \in D_k\}$ based on these neighbors’ integrated scores. The conditional probability of the test gene $g_t$ with respect to disease (+)/non-disease class (−), is measured as

$$P(+|g_t,h=WKN) = \frac{\sum_{g \in D_p} [\text{Int}_\text{score}(g)]}{\sum_{g \in D_k} [\text{Int}_\text{score}(g)]}$$

$$P(−|g_t,h=WKN) = \frac{\sum_{g \in D_0} [\text{Int}_\text{score}(g)]}{\sum_{g \in D_k} [\text{Int}_\text{score}(g)]}$$

(7)

Note that weighted KNN accumulates both positive and negative integrated scores in $D_k$ and estimates the probability of $g_t$ belonging to positive (or negative) class based on the accumulated scores in that class.

**PU learning model 2: Weighted Naive Bayes (WNB).** Given a test gene $g_t$, the probability that gene $g_t$ belongs to a class $c_j$ ($c_j \in C = \{+,−\}$) can be computed using Bayes’ theorem as:

$$P(Y = c_j|g_t,h=WNB) = \frac{P(g_t|Y = c_j)P(Y = c_j)}{P(g_t)}$$

(8)

where the probability $P(g_t)$ is a constant for the positive and negative classes. Here, we define the prior probabilities of the positive and negative classes as 0.5, i.e. $P(Y=+)=P(Y=−)=0.5$. Given a gene vector $\vec{g} = \{g_1,\ldots,g_m\}$, the conditional probability of feature $f_k$ associated with class $c_j$, denoted as $P(f_k|Y=c)$, is calculated as:

$$P(f_k|Y=c) = \sum_{g \in D_k} g(f_k) * \text{Int}_\text{score}(g)$$

(9)

where $g(f_k)$ is the value of feature $f_k$ in gene vector $\vec{g}$, $D_k$ is defined as either $D_p = \{g | \text{Int}_\text{score}(g) > 0\}$ or $D_0 = \{g | \text{Int}_\text{score}(g) < 0\}$, depending on $c_j$ is positive class (+) or negative class (−).

By assuming that the probabilities of features are independent given the class $Y=c_j$, we obtain the Naive Bayes classifier:

$$P(Y = c_j|g_t,h=WNB) = \frac{\prod_{k=1}^{m} g(f_k|Y=c_j)P(f_k|Y=c_j)}{\prod_{k=1}^{m} P(f_k|Y=c_j)}$$

(10)

**PU learning model 3: Multi-level Support Vector Machine (MSVM).** Based on the integrated score $\text{Int}\text{score}(g)$, we further partition the unlabeled genes $g \in (U-RN)$ into three parts: likely positive set $LP$ (genes get higher positive integrated scores), likely negative set $LN$ (genes get lower negative integrated scores) and weak negative set $WN$ (remaining genes) using the following criteria:

$$L(g) = \begin{cases} 
LP & \text{Int}_\text{score}(g) > (1-\alpha) \\
LN & \text{Int}_\text{score}(g) < -(1-\alpha) \\
WN & -(1-\alpha) \leq \text{Int}_\text{score}(g) \leq (1-\alpha)
\end{cases}$$

(11)

We then build a multi-level classifier based on positive training set $P$, reliable negative set $RN$, and three newly generated sets $LP$, $LN$, and $WN$, via weighted support vector machine technique [28][29], to take into account of the inherently different levels of trustworthiness of labels in the five gene set.

The objective function of Weighted SVM can be defined as [14]:

$$\text{Minimize: } \frac{1}{2}||w||^2 + c_+ \sum_{i \in LP} \xi_i + c_- \sum_{i \in LN} \xi_i + c_w \sum_{i \in WN} \xi_i$$

(12)

Subject to: $y_i(w^T x_i + b) \geq 1 - \xi_i (i=1,2,\ldots,m)$ where the values of parameters $c_+, c_-, c_w, c_w$ and $c_-\in [1,2,\ldots,\infty)$ can be decided by using cross-validation techniques. Finally, we apply our MSVM model $P(Y = c_j|g_t,h=MSTM)$ to compute the probability of test gene $g_t$ belonging to class $c_j$ ($c_j \in C = \{+,−\}$) for its classification.

Note that while the candidate positive set $CP$ plays a role in assigning the genes in $U - RN$ to one of the 3 subsets $LP$, $LN$ and $WN$, it does not overlap with the training set $P$, $U$ and hence is not used in the construction of the MSTM model.

**Ensemble-based algorithm for integration of individual classifiers.** In order to perform more robust classification, we design a novel ensemble learning model to integrate the three classification models constructed above.

Suppose $x_{ij} = P(Y = c_j|g_t,h_j)$ denotes the probability of gene $g_t$ belonging to class $c$ as predicted by the $j$th classifier. We can organize the genes in $D$ in the following matrix:

$$\begin{bmatrix}
x_{11} & \cdots & x_{1k} \\
\vdots & \ddots & \vdots \\
x_{m1} & \cdots & x_{mk}
\end{bmatrix}$$

(13)

where $k$ is the number of individual classifiers (here, $k=3$), and $|D|$ is the size of training set $D$.

Our ensemble model $o(\vec{x})$ integrates the outputs from the multiple classification models as follows:
\[ o(x_1, \ldots, x_d) = \text{sgn}(\vec{w} \cdot \vec{x}_i) \]
\[ = \begin{cases} 
1, & \text{if } w_0 + w_1 x_{i1} + \ldots + w_k x_{ik} > 0 \\
-1, & \text{otherwise}
\end{cases} \quad (14) \]

where \( \vec{w} \) is a weight vector that indicates the importance of individual models. The final output value "1" denotes disease/positive class and '-1' denotes non-disease/negative class.

The classifier weight \( \vec{w} \) can be learned from training set \( D \) as follows. We define \( E(\vec{w}) \) as training error of the hypothesis of our ensemble model:

\[ E(\vec{w}) = \frac{1}{2} \sum_{i \in D} (y_i - o_i)^2 \quad (15) \]

where \( y_i \), \( o_i \in \{-1, 1\} \) and \( o_i \in \{-1, 1\} \) are the actual class and predicted class by our ensemble model for training gene \( g_i \) respectively. \( E(\vec{w}) \) is a linear square error function that evaluates the difference between \( y_i \) and \( o_i \). We minimize \( E(\vec{w}) \) to guarantee the classification output \( o \) with minimal error rate and calculate the weight vector \( \vec{w} \).

Here, gradient decent is applied to search the probable weight vectors in error surface. The gradient of \( E(\vec{w}) \) for \( \vec{w} \), denoted as \( \nabla E(\vec{w}) = \left[ \frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \ldots, \frac{\partial E}{\partial w_k} \right] \), is the derivative of \( E \) with respect to each component of the vector \( \vec{w} \). From above equation, we obtain each component of \( \nabla E(\vec{w}) \) as follows:

\[ \frac{\partial E}{\partial w_j} = \frac{1}{2} \sum_{i \in D} \frac{\partial}{\partial w_j} (y_i - o_i)^2 = \frac{1}{2} \sum_{i \in D} \frac{\partial}{\partial w_j} (y_i - o_i)^2 \]
\[ = \sum_{i \in D} (y_i - o_i) \frac{\partial}{\partial w_j} (y_i - \vec{w} \cdot \vec{x}_i) \]
\[ = \sum_{i \in D} (y_i - o_i) (-x_{ij}) \quad (16) \]

The following training rule guarantees that \( \vec{w} \) is adjusted in the direction of steepest descent along the error surface: \( \vec{w} \leftarrow \vec{w} + \Delta \vec{w} \), where \( \Delta \vec{w} = -\eta \nabla E(\vec{w}) \). \( \eta \) is a small positive constant, called learning rate, to determine the step size in gradient decent exploration. We set \( \eta = 0.001 \), following previous work [30]. The negative gradient \( -\nabla E(\vec{w}) \) gives the direction of steepest decrease. According to equations above, we update the gradient descent rule as follows:

\[ \Delta w_j = -\eta \frac{\partial E}{\partial w_j} = \eta \sum_{i \in D} (y_i - o_i) x_{ij} \quad (17) \]

The overall ensemble learning method is summarized in Figure 2. First, we assign an initial random weight vector \( \vec{w} \). The ensemble model is then applied to all training genes and each weight is then updated by adding \( \Delta w_j \) computed according to equation (17) above. This process is repeated until \( \vec{w} \) converges.

### Ensemble learning algorithm

1. Initialize each element in weight vector \( \vec{w} \) with a small random real number;
2. Do following operations until \( \vec{w} \) converges:
   3. Initialize each \( \Delta w_j = 0 \);
   4. For each decision vector \( <x_i, y> \) in training samples, Do:
      5. compute output \( o(x_i) = \text{sgn}(\vec{w} \cdot \vec{x}_i) \);
   6. For each linear unit weight \( w_j \), Do:
      7. \( \Delta w_j = \Delta w_j + \eta (y - o) x_{ij} \)
   9. \( w_j = w_j + \Delta w_j \)
10. Record the weight vector as optimal vector;
11. Predict the class of test samples using \( o(x_1, \ldots, x_k) = \text{sgn}(\vec{w} \cdot \vec{x}_i) \)

### Experimental Results

For evaluation, we benchmark our proposed EPU algorithm against four state-of-the-art techniques for disease gene prediction: PUDI method, Smalter’s method, Xu’s method and ProDiGe. In addition, we also compare the performance of EPU with its base learning models, namely MSVM, WKNN and WNB. Finally, we demonstrate novel disease gene prediction using the EPU algorithm.

**Experimental settings**

We use the disease classes with at least 50 confirmed disease genes from the 21 specific disease classes in [6] for evaluating our classification algorithm. There are six such disease classes: cardiovascular disease, endocrine disease, cancer disease, metabolic disease, neurological disease, and ophthalmological disease (See Table S1 for the exact numbers of disease genes for each class). Given a particular disease class, the positive set \( P \) consists of all its confirmed disease genes, while the unlabeled set \( U \) is formed by randomly selecting from the genes that are not known to be associated with any disease such that \( |P| = |U| \), following the setting in [9] [10] [15]. To avoid bias in sampling, we randomly select 10 groups of unlabeled set \( U \). All experimental evaluations of the classification models are done on identical groups of training and test data, and we report the average performance over the 10 groups of \( P+U \) sets. To evaluate the performance of our algorithm, 3 fold cross validation is applied where two folds in \( P+U \) as the training set build classifier while remaining one fold is the test set. Next, positive training genes in \( P \) are used as seed nodes on multiple genetic networks to weight unlabeled training genes in \( U \) via flow propagation. Then, to obtain the ‘soft’ classes of training genes from component learning models, leave-one-out cross-validation (LOOCV) is used on 2 fold training samples, from which each training sample is singled out to evaluate its ‘soft’
classes by component classifiers (MSVM, WKNN and WNB), which are built on the other training samples. After LOOCV, these training samples with the ‘soft’ classes are as input data to build an ensemble learning model (as shown in Figure 2) to predict 1 fold test fold. The average results on 3 \times 10 groups of \( (P+U) \) sets are reported on experimental part.

### Evaluation metrics

We use precision, recall, and F-measure to measure the performance of our classification models on each of the six disease classes. The F-measure is the harmonic mean of precision (denoted as \( p \)) and recall (denoted as \( r \)), defined as

\[
F = \frac{2pr}{p+r}
\]

(18)

The value of F-measure is large only when both of \( p \) and \( r \) are high, and small when either of them is poor. This is appropriate for our objective of accurately predicting disease genes, as deficiencies in either precision or recall will be reflected by a low F-measure.

<table>
<thead>
<tr>
<th>Disease group</th>
<th>Techniques</th>
<th>Precision (p)</th>
<th>Recall (r)</th>
<th>F-measure (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular</td>
<td>PUDI</td>
<td>82.0%</td>
<td>80.3%</td>
<td>80.4%</td>
</tr>
<tr>
<td></td>
<td>ProDiGe</td>
<td>54.3%</td>
<td>96.3%</td>
<td>69.3%</td>
</tr>
<tr>
<td></td>
<td>Smalter’s method</td>
<td>75.4%</td>
<td>67.6%</td>
<td>70.6%</td>
</tr>
<tr>
<td></td>
<td>Xu’s method</td>
<td>72.1%</td>
<td>60.0%</td>
<td>65.4%</td>
</tr>
<tr>
<td></td>
<td>EPU</td>
<td>85.2%</td>
<td>81.0%</td>
<td>84.1%</td>
</tr>
<tr>
<td>Endocrine</td>
<td>PUDI</td>
<td>83.6%</td>
<td>75.3%</td>
<td>79.2%</td>
</tr>
<tr>
<td></td>
<td>ProDiGe</td>
<td>57.3%</td>
<td>87.7%</td>
<td>69.3%</td>
</tr>
<tr>
<td></td>
<td>Smalter’s method</td>
<td>76.4%</td>
<td>58.8%</td>
<td>66.5%</td>
</tr>
<tr>
<td></td>
<td>Xu’s method</td>
<td>75.4%</td>
<td>62.0%</td>
<td>68.0%</td>
</tr>
<tr>
<td></td>
<td>EPU</td>
<td>88.1%</td>
<td>87.7%</td>
<td>87.9%</td>
</tr>
<tr>
<td>Neurological</td>
<td>PUDI</td>
<td>70.3%</td>
<td>80.1%</td>
<td>74.9%</td>
</tr>
<tr>
<td></td>
<td>ProDiGe</td>
<td>63.1%</td>
<td>74.0%</td>
<td>68.1%</td>
</tr>
<tr>
<td></td>
<td>Smalter’s method</td>
<td>60.6%</td>
<td>65.9%</td>
<td>63.1%</td>
</tr>
<tr>
<td></td>
<td>Xu’s method</td>
<td>59.7%</td>
<td>66.7%</td>
<td>63.0%</td>
</tr>
<tr>
<td></td>
<td>EPU</td>
<td>78.2%</td>
<td>80.4%</td>
<td>78.6%</td>
</tr>
<tr>
<td>Metabolic</td>
<td>PUDI</td>
<td>80.1%</td>
<td>84.8%</td>
<td>82.4%</td>
</tr>
<tr>
<td></td>
<td>ProDiGe</td>
<td>58.7%</td>
<td>84.5%</td>
<td>69.3%</td>
</tr>
<tr>
<td></td>
<td>Smalter’s method</td>
<td>59.1%</td>
<td>84.7%</td>
<td>69.6%</td>
</tr>
<tr>
<td></td>
<td>Xu’s method</td>
<td>65.6%</td>
<td>78.3%</td>
<td>71.4%</td>
</tr>
<tr>
<td></td>
<td>EPU</td>
<td>83.3%</td>
<td>93.9%</td>
<td>90.9%</td>
</tr>
<tr>
<td>Ophthalmological</td>
<td>PUDI</td>
<td>71.6%</td>
<td>78.5%</td>
<td>74.9%</td>
</tr>
<tr>
<td></td>
<td>ProDiGe</td>
<td>58.3%</td>
<td>77.7%</td>
<td>66.6%</td>
</tr>
<tr>
<td></td>
<td>Smalter’s method</td>
<td>56.7%</td>
<td>77.8%</td>
<td>65.5%</td>
</tr>
<tr>
<td></td>
<td>Xu’s method</td>
<td>64.2%</td>
<td>71.3%</td>
<td>67.4%</td>
</tr>
<tr>
<td></td>
<td>EPU</td>
<td>89.3%</td>
<td>81.0%</td>
<td>84.7%</td>
</tr>
<tr>
<td>Cancer</td>
<td>PUDI</td>
<td>76.3%</td>
<td>80.0%</td>
<td>78.0%</td>
</tr>
<tr>
<td></td>
<td>ProDiGe</td>
<td>71.1%</td>
<td>79.8%</td>
<td>75.3%</td>
</tr>
<tr>
<td></td>
<td>Smalter’s method</td>
<td>73.8%</td>
<td>79.0%</td>
<td>76.3%</td>
</tr>
<tr>
<td></td>
<td>Xu’s method</td>
<td>71.0%</td>
<td>79.7%</td>
<td>75.1%</td>
</tr>
<tr>
<td></td>
<td>EPU</td>
<td>81.2%</td>
<td>84.5%</td>
<td>82.6%</td>
</tr>
<tr>
<td>Average performance</td>
<td>PUDI</td>
<td>77.3%</td>
<td>79.8%</td>
<td>78.3%</td>
</tr>
<tr>
<td></td>
<td>ProDiGe</td>
<td>60.5%</td>
<td>83.3%</td>
<td>69.7%</td>
</tr>
<tr>
<td></td>
<td>Smalter’s method</td>
<td>67.0%</td>
<td>72.3%</td>
<td>68.6%</td>
</tr>
<tr>
<td></td>
<td>Xu’s method</td>
<td>68.0%</td>
<td>69.7%</td>
<td>68.4%</td>
</tr>
<tr>
<td></td>
<td>EPU</td>
<td>84.2%</td>
<td>84.8%</td>
<td>84.8%</td>
</tr>
</tbody>
</table>
Experimental Results

Benchmarking of EPU ensemble learning algorithm against state-of-art techniques. First, we compared our EPU algorithm against four state-of-the-art techniques, namely, PUDI method [13], Smalter’s method [10], Xu’s method [15] and ProDiGe [12]. Table 1 shows that our proposed EPU, on average, is 6.5%, 15.1%, 16.2% and 16.4% better than PUDI, ProDiGe, Smalter’s method, Xu’s method in terms of F-measure respectively. In particular, EPU can achieve much better precision and consistently better recall when compared against the recently proposed method PUDI. It shows that EPU can effectively extract hidden positive and negative data from the unlabeled data to boost classification performance.

Comparison of EPU with base classifiers. Next, we compared the performance of our proposed EPU against its base classifiers MSVM, WNB, WKNN. As shown in Table 2, on average, MSVM achieved the highest F-measure (81.3%), much higher than WNB (69.5%) and WKNN (68.7%). This is not surprising as MSVM can handle multiple weighted positive and negative sets when building its classification model. Furthermore, SVM is known to perform significantly better than NB and KNN in many real-world applications.

Our proposed ensemble learning method EPU is able to achieve 84.8% in terms of F-measure, which is 3.5%, 15.3% and 16.1% better than MSVM, WNB and WKNN respectively. Moreover, EPU consistently outperformed all 3 component classifiers for every disease class. This strongly demonstrates that EPU can effectively integrate multiple classification models and minimize the overall error rate through dynamically assigning different weights to different classification models.

Sensitivity study on the parameter $g$ in EPU and parameter $k$ in genetic similarity networks. We perform the sensitivity study for parameter $g$ in EPU and coverage of genetic similarity networks. Parameter $g$ is the learning rate in EPU algorithm. We perform EPU on six disease groups with $g$ from 0.001 to 0.03. The result indicates that step size within 0.001 is small enough to move optimal value point in hypothesis space and our EPU is robust and stable when $g$ is small (Table S3 for the detailed results). In addition, we study the effect of the parameter $k$ that determines the number of neighbors of each gene in biological networks. The results in Table S4 show that EPU consistently achieved best performance with $k$ in (1, 9).

Comparing EPU with existing ensemble learning approaches. We also compared our proposed EPU with two

<table>
<thead>
<tr>
<th>Disease group</th>
<th>Techniques</th>
<th>Precision ($\rho$)</th>
<th>Recall ($r$)</th>
<th>F-measure ($F$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular</td>
<td>MSVM</td>
<td>74.3%</td>
<td>87.6%</td>
<td>80.4%</td>
</tr>
<tr>
<td></td>
<td>WNB</td>
<td>57.3%</td>
<td>72.5%</td>
<td>63.9%</td>
</tr>
<tr>
<td></td>
<td>WKNN(3)</td>
<td>60.1%</td>
<td>68.6%</td>
<td>64.0%</td>
</tr>
<tr>
<td></td>
<td>EPU</td>
<td>85.2%</td>
<td>81.0%</td>
<td>84.1%</td>
</tr>
<tr>
<td>Endocrine</td>
<td>MSVM</td>
<td>83.4%</td>
<td>85.2%</td>
<td>84.2%</td>
</tr>
<tr>
<td></td>
<td>WNB</td>
<td>61.3%</td>
<td>70.4%</td>
<td>65.3%</td>
</tr>
<tr>
<td></td>
<td>WKNN(3)</td>
<td>64.5%</td>
<td>53.1%</td>
<td>57.9%</td>
</tr>
<tr>
<td></td>
<td>EPU</td>
<td>88.1%</td>
<td>87.7%</td>
<td>87.9%</td>
</tr>
<tr>
<td>Neurological</td>
<td>MSVM</td>
<td>69.3%</td>
<td>83.7%</td>
<td>75.8%</td>
</tr>
<tr>
<td></td>
<td>WNB</td>
<td>61.1%</td>
<td>74.4%</td>
<td>67.0%</td>
</tr>
<tr>
<td></td>
<td>WKNN(3)</td>
<td>62.3%</td>
<td>67.1%</td>
<td>64.6%</td>
</tr>
<tr>
<td></td>
<td>EPU</td>
<td>78.2%</td>
<td>80.4%</td>
<td>78.6%</td>
</tr>
<tr>
<td>Metabolic</td>
<td>MSVM</td>
<td>84.0%</td>
<td>91.3%</td>
<td>87.4%</td>
</tr>
<tr>
<td></td>
<td>WNB</td>
<td>68.8%</td>
<td>79.9%</td>
<td>73.9%</td>
</tr>
<tr>
<td></td>
<td>WKNN(3)</td>
<td>76.6%</td>
<td>78.8%</td>
<td>77.6%</td>
</tr>
<tr>
<td></td>
<td>EPU</td>
<td>83.3%</td>
<td>93.9%</td>
<td>90.9%</td>
</tr>
<tr>
<td>Ophthalmological</td>
<td>MSVM</td>
<td>78.4%</td>
<td>86.1%</td>
<td>81.9%</td>
</tr>
<tr>
<td></td>
<td>WNB</td>
<td>61.2%</td>
<td>78.7%</td>
<td>68.8%</td>
</tr>
<tr>
<td></td>
<td>WKNN(3)</td>
<td>67.3%</td>
<td>72.2%</td>
<td>69.6%</td>
</tr>
<tr>
<td></td>
<td>EPU</td>
<td>89.3%</td>
<td>81.0%</td>
<td>84.7%</td>
</tr>
<tr>
<td>Cancer</td>
<td>MSVM</td>
<td>73.4%</td>
<td>83.9%</td>
<td>78.3%</td>
</tr>
<tr>
<td></td>
<td>WNB</td>
<td>72.5%</td>
<td>85.1%</td>
<td>78.3%</td>
</tr>
<tr>
<td></td>
<td>WKNN(3)</td>
<td>76.4%</td>
<td>81.0%</td>
<td>78.6%</td>
</tr>
<tr>
<td></td>
<td>EPU</td>
<td>81.2%</td>
<td>84.5%</td>
<td>82.6%</td>
</tr>
<tr>
<td>Average performance</td>
<td>MSVM</td>
<td>78.6%</td>
<td>86.3%</td>
<td>81.3%</td>
</tr>
<tr>
<td></td>
<td>WNB</td>
<td>63.7%</td>
<td>76.8%</td>
<td>69.5%</td>
</tr>
<tr>
<td></td>
<td>WKNN(3)</td>
<td>67.9%</td>
<td>70.1%</td>
<td>68.7%</td>
</tr>
<tr>
<td></td>
<td>EPU</td>
<td>84.2%</td>
<td>84.8%</td>
<td>84.8%</td>
</tr>
</tbody>
</table>

Table 2. Overall comparison to single-expert classifiers.

EPU is compared with its three component classifiers Multi-level Support Vector Machine (MSVM), Weighted Naïve Bayes (WNB) and Weighted K-Nearest Neighbor (KNN) on 6 disease groups. WKNN(3) is an instance-based classifier that predicts the class of an unlabeled gene based on its 3 closest labeled genes.

doi:10.1371/journal.pone.0097079.t002
existing ensemble approaches, majority vote and weighted majority vote [31], in terms of F-measure across the six disease classes. EPU was shown to outperform the existing ensemble methods (see Table S2 for the detailed results), indicating that our proposed EPU is a superior ensemble strategy for integrating multiple classification models for disease gene prediction.

**Predicting novel disease genes for disease groups.** To demonstrate novel disease gene prediction using the EPU algorithm, we selected two important disease groups, namely, metabolic and cancer, as detailed case studies. For each target disease class, we obtained a set of confirmed disease genes from OMIM and GENECARD as the positive training set, and applied our proposed EPU algorithm to prioritize a novel disease gene from the unlabeled gene set.

We first applied our EPU algorithm to discover novel disease genes for metabolic diseases. 12 unlabeled genes were detected to be associated with target disease using our algorithm. For verification, we searched the literature for evidence that supports the association of these predicted disease genes to metabolic diseases. We found that two predicted genes, RHEB and DOK5, have indeed been reported to be associated with metabolic diseases. RHEB, a GTP-binding protein, was reported to be inactivated to protect cardiomyocyte during energy deprivation via activation of autophagy. This implies that RHEB is a key regulator of autophagy during myocardial ischemia, which has implications in patients with obesity and metabolic syndrome [32]. As for DOK5, Tabassum et al. identified that it is a novel candidate disease genes associated with type 2 diabetes, which is a metabolic disorder due to obesity [33].

Our EPU model also predicted 32 unlabeled genes as candidate genes associated with cancer. Seven of them, SIGLEC7, PRDX4, PRDX5, HNRNPL, SRPK1, ABCB10 and PHF10 have been reported to be associated with cancer diseases. Table 3 lists these candidate disease genes and the supporting literature evidence that we have found.

For other candidate cancer genes without literature evidence support, seven of them, PMM1, SRCIN1, ISY1, KDM4A, CIR1, PPP2R5A and NOL3 have been shown to associate with cancer diseases in GO similarity network, GE similarity network and PPI network. From GO similarity network, PMM1 is one of top 5 nearest neighbours of cancer disease gene PPM1D and SRCIN1 is one of neighbours of disease gene CTNNB1. In GE similarity network, ISY1 is linked to disease gene P2RX7, KDM4A and CIR1 are interacted with disease genes CTNNB1 and MSH2 respectively, indicating that three suspicious genes are highly correlated with cancer disease genes in terms of gene expression. From PPI network, PPP2R5A is directly interacted with two disease genes, BCL2 and TP53, and NOL3 is linking to two disease genes, BAX and CASP8.

**Conclusions and Discussion**

Despite the considerable progress in disease gene discovery, there are still many unknown disease genes that are yet to be characterized. Machine learning methods can be used to predict novel disease genes from the confirmed disease genes, based on the observation that genes associated with similar disease phenotypes are likely to share similar biological characteristics. However, there are two challenging issues for disease gene predictions. Firstly, how to leverage various biological sources during our model building process, which could effectively alleviate the bias issues from the incompleteness and noise in the data. Secondly, how to integrate...
In this paper, we choose Nearest Neighbor, Naïve Bayes and SVM potential bias, incompleteness and noise in individual data source. using multiple biological data sources, EPU is less susceptible to multiple biological networks, and an ensemble classification approach on multiple machine-learned prediction models. By using multiple biological data sources, EPU is less susceptible to potential bias, incompleteness and noise in individual data source. In this paper, we choose Nearest Neighbor, Naïve Bayes and SVM as three base learning models of EPU due to three reasons: they are the state-of-the-art learning techniques that have been widely used in disease gene identification [10] [11] [15] [17] [34]; 2) we are combining PU learning models instead of traditional classification models – we choose the three classification models as they can be easily adapted to build PU learning models; 3) they are quite diverse with learning criterions, so that their complementary nature may contribute a more accurate and robust combinatorial result. By employing an ensemble approach for prediction, EPU also minimizes the inherent limitations of individual prediction models. Finally, by employing PU learning techniques for building its ensemble of classification models, EPU is able to treat the unknown genes appropriately as an unlabeled set \( U \) (instead of a negative set \( N \)) for training, thereby resulting in more robust predictions. Experimental evaluations have confirmed the effectiveness of our proposed approach, with our EPU method consistently performing much better than the existing state-of-the-art techniques for disease gene prediction on six disease classes.

As more biological data sources and machine learning classifiers become available in the future, our EPU method can be an effective framework to integrate the additional biological and computational resources for better disease gene predictions. For further work, we will explore the inclusion of other biological data sources for disease gene prediction using our framework. Given that many machine learning problems in biomedical research do involve ensemble Positive Unlabeled data, we can also adapt our EPU framework to other applications, such as drug-target interaction prediction [35] [36].

Supporting Information

Figure S1 Procedure of extracting candidate positive set. (DOCX)
Table S1 Number of disease genes associated with six disease classes. (DOCX)
Table S2 Performance comparison of ensemble methods. (DOCX)
Table S3 Effect of parameter \( \eta \) on classification performance of six disease groups. (DOCX)
Table S4 Sensitive analysis on biological network noise to disease gene prediction. (DOCX)

Author Contributions

Conceived and designed the experiments: PY XL.Performed the experiments: PY. Analyzed the data: PY H-NC. Contributed reagents/materials/analysis tools: H-NC. Wrote the paper: PY XL. Participated in discussion and improved the method as well as revised the draft: C-KK S-KN. Read and approved the manuscript: PY XL H-NC C-KK S-KN.

References


