ORACLE CHARACTERIZATION FOR ACTIVE LEARNING FOR PROTEIN-PROTEIN INTERACTION PREDICTION

Seshan Ananthasubramanian\textsuperscript{1,2}, Jaime G. Carbonell Madhavi\textsuperscript{3} and K. Ganapathiraju\textsuperscript{1,2,3}

\textsuperscript{1}Department of Biomedical Informatics and \textsuperscript{2}Intelligent Systems program, University of Pittsburgh, 5607 Baum Blvd, Suite 401, Pittsburgh, PA, 15206, USA, \textsuperscript{3}Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA, madhavi@pitt.edu, madhavi@cs.cmu.edu

ABSTRACT

Discovery of human interactome is crucial for the understanding of complex biological processes and pathways, and for drug discovery. Bench-work experiments to determine protein-protein interactions (PPIs) are costly in terms of time, materials, equipment, and technical and scientific expertise. Thus, in recent years computational machine learning methods have been proposed to predict PPIs. These methods require training data to learn the classification models; but, currently known interactions are not sufficient to create an accurate model to predict the whole interactome or even a significant fraction thereof. If it is possible to seek additional training data experimentally, which are the protein-pairs that should be chosen to yield maximally-informative instances so as to maximize prediction accuracy for a given cost/effort level? Active machine learning (AL) methods are designed to select instances (protein-pairs) to create optimal training data and thus maximize their value in terms of prediction accuracy. Improving the relative accuracy is the pursuit of research in AL. However AL assumes the existence of an omniscient oracle (an experiment or an expert), which would give the correct label for every instance that it is asked, and it would do so for both positive and negative labels. In reality in the context of PPIs, an “oracle” can provide labels for only some interacting pairs (reductance), and cannot give labels for non-interacting pairs (one-class nature). We develop algorithms for AL appropriate to this scenario of PPI prediction. Our results are superior compared to both a random baseline and also generic state-of-the-art AL.

1. INTRODUCTION

Computationally discovering protein-protein interactions (PPIs) of the human interactome is a challenging task for many reasons. Supervised machine learning algorithms have been applied to PPI prediction, which treat the task as a binary classification problem. Positive class data, namely the interacting protein-pairs, are only one-in-a-thousand or less from amongst 500 million possible pairs [1]. There are no known negative-class data, namely virtually no probably non-interacting pairs. Therefore, a classification model must be built using training data that consists of a pool of known interactions and a pool of randomly paired proteins not known to interact that are treated as non-interacting pairs [2]. The best estimates of the total number of interactions [3] indicate that only 5-10% are currently known, and it is possible that these interactions are not representative of the entire interaction space – i.e. they are not drawn randomly or i.i.d. from the interacting distribution. For more than half of all the proteins, there is not even one known interaction. In order to learn an accurate PPI classifier, one must sample the space and determine labels of those instances by experimental methods. However, acquiring more interactions would require performing bench work experiments, including medium to high-throughput methods such as yeast 2-hybrid [4] or mass spectrometry techniques [5], which are costly and time consuming processes that also require considerable amount of resources, high-end equipment and technical expertise. Therefore, designing strategies that optimally select instance pairs that are most informative for incrementally training an accurate classifier is an important goal in PPI prediction. This is known as active learning.

Active learning strategies help to select optimal training instances to achieve superior predictive accuracy within a given budget that is available for labeling instances [6]. See Figure 1. In active learning, the algorithm starts with the few labeled instances that are available (‘\(\bullet\)’ and ‘\(\circ\)’ in Figure 1). Next, it identifies the unlabeled instances (pairs of proteins) whose labels when known would prove most useful in learning a better classifier (‘\(\ast\)’ in Figure 1); an oracle is queried to obtain labels of those instances; the labels thus obtained are added to the training data and the model is re-trained to arrive at a more accurate classifier; if time and budget permit, this process is iterated. The basic assumption is that obtaining labels involves investment of resources, and therefore that the selected instances whose labels are asked should be optimal for retraining more accurate classifier. In PPI prediction, an oracle would typically be a bench-work experiment, which can characterize a protein-pair. Common strategies for active learning include density-based selection where more representative instances are selected from denser clusters [7], or uncertainty-based selection where data points are selected from maximum
confusion or uncertain regions of the instance space with respect to the current classifier [8], or ensemble-methods which employ multiple criteria to select data points that typically outperform many other strategies [9]. “Wrapper” methods select those instances which would lead to the highest improvement in classifier accuracy once they are added to the training dataset [6] by the computationally-intensive step of hypothesizing the label of each instance, retraining the classifier as if this hypothesis were true, and estimating accuracy gains (for PPI this would require a half-billion retraining steps to select a single instance for experimentation; hence it is not a tractable option). We previously applied a variety of active learning methods including density based, uncertainty based and history based active selection approaches for predicting PPIs [10] in which we observed that active learning required only 500 labeled instances to achieve the same or superior accuracy as achieved by 3,000 randomly selected labeled instances. A six-fold improvement in experimental efficiency with modest computational effort is always a desirable tradeoff, but can we do better? This paper argues the positive.

2. APPROACH

Active learning assumes that there exists a single perfect oracle, which would always give the correct answer for labeling instances — e.g. an experimental procedure that always yields an answer and is always correct — this is unrealistic of experiments to determine protein-protein interactions. These characteristics of the oracle present a hindrance to optimal active learning for interactome-scale discovery of PPIs. Would active learning work better with a proper “oracle” for PPI? This is the central hypothesis of this paper, which we answer in the positive. The primary contributions of this work are (a) a detailed characterization of the oracle in the domain of PPI prediction, and (b) development of suitable active learning approaches to suit these oracle characteristics, and an empirical demonstration of their effectiveness over the state of the art.

2.1. Characteristics of the PPI oracle

The PPI oracle is not able to detect non-interacting proteins (one-class nature): Traditional bench-work experiments cannot validate “non-interaction nature” of protein pairs. There is no data available which gives us information about experimentally confirmed non-interacting proteins. Thus, PPI oracles can only provide labels for protein-pairs associated with a single class, which is the interacting class.

The PPI oracle is only able to label a subset of all real interactions (reluctance): There are various technological constraints which limit the experiment to detect all possible interactions. Some experiments cannot be used on a certain class of proteins. For example, a yeast 2-hybrid (Y2H) experiment cannot be used for detecting interactions of a protein that is able to initiate transcription without its interacting partner or those that
are dependent on post-translational modifications; mass-spectrometry methods might fail to discover transient interactions; most experimental techniques that are available today are not able to characterize interactions involving integral membrane proteins [11]. Thus, technical constraints constitute a major roadblock for many experimental methods, and they are only able to identify a subset of all possible interactions.

![Cumulative Interactions](image)

**Figure 2 – Number of known interactions of human genes:** Graph showing percentage of known interactions against the total number of genes associated with those interactions, sorted in descending order of the number of interactions. 50% known interactions are associated with 4% of genes.

**The PPI oracle labels a subset of interactions incorrectly (fallibility):** Human errors and operation issues while performing the experiment can lead to incorrect labeling of some protein-pairs. A two hybrid assay can produce some biologically irrelevant interactions, especially if the proteins reside in different tissues or different sub cellular locations [11]. It is also difficult to isolate the binary interactions between protein-pairs when a protein complex is involved, as it is highly difficult to identify the target protein in the complex. Many high-throughput interactions are detected in-vivo by causing disruption of normal cellular function [11]. Thus non-typical interactions may be observed, as the existing pattern of protein interactions is disrupted resulting in the generation of false positives.

2.2. **Active learning with PPI oracles**

We extend general active learning methods to the scenario where the oracle has one class and reluctance properties, considering these characteristics one at a time. We consider the oracle to be intractable in nature; that is, if it does give a label, it is believed to be the correct label. We call this Active Learning with Reluctant One Class oracle. Whereas the reluctance property and the accuracy of oracles were introduced and characterized by Donmez et al [12], the one-class property is a new contribution to active learning driven by the requirements of PPI prediction.

2.2.1. **Active learning with one class oracle**

In classical active learning, if an oracle is resultant (no answer), we assume the majority class, i.e. the protein pair is non-interacting. However in one class active learning, if the oracle fails to give the label for a data point, then we estimate whether a potential label can be associated with it by observing the oracle behavior. We assign the estimated label to the data point and add it to the training data instead of always assuming it to be non-interacting in nature. We also estimate \( P(label \mid x) \), which is the probability that the data-point \( x \) would be labeled by the oracle. As the oracle provides the label for only the interacting class, we can use \( P(label \mid x) \) as a reasonable estimate to assign the “interacting” label to the instance. The unlabeled instance is assigned the interacting label with a probability \( P(label \mid x) \).

There exists no real world datasets from which we can learn the behavior of the PPI oracle to calculate \( P(label \mid x) \) for unlabeled instances. Hence, we propose three different heuristic methods which provide a reasonable estimate of \( P(label \mid x) \) based on the distribution of known labeled data that have already been added to the training data in the active learning process. These heuristic methods are described in the next section. The following are assumptions in determining \( P(label \mid x) \) through heuristic based methods:

a. We assume that if the oracle gives the (interacting) label for a particular data-point, then it is highly likely to assign the same label for neighboring points in PP feature space. As the distance from known labeled instances increases, the likelihood of an instance being labeled by the oracle decreases.

b. Some proteins have been studied extensively due to their significant role in a significant pathway, or their role as a drug target or due to disease-association. For 9673 genes out of 22,500 genes there is at least one known interaction (Human Protein Reference Database: www.HPRD.org [13]); 60% of all genes have no known experimentally determined interaction. Figure 2 shows number of interactions of proteins ordered by their rank when ordered descending by the number of interactions known. It can be seen from the figure that around 50% of all known interactions are associated with only 1,250 genes or 4% of all genes. Thus most known interactions today are associated with a very few proteins. The labeled subset space is generally associated with those 4% of genes that are well studied in literature.

2.3. **Active learning with reluctant oracle**

When an oracle is asked for labels of unlabeled instances \( U \), of which \( N \) are interacting pairs in reality, the oracle gives labels for only \( n \) out of \( N \) interacting pairs owing to its reluctance nature described earlier. For example, in a yeast 2-hybrid set up, the “assay sensitivity” was found to be only 23% [14]. In order that all unlabeled interactions be recovered by the learning algorithm, it is expected that these \( n \) labels be a random subset from the \( N \) interactions for which the label was asked. If the
labeled positive examples are indeed a random subset of all positive examples for which labels are asked, then the true conditional probability that an instance is interacting, given by \( P(y = \text{"interacting"} | x) \) and probability that the instance is labeled, given by \( P(\text{label} | x) \), differ only by a constant factor [15]. Thus,
\[
P(y = \text{"interacting"} | x) = c \ast P(\text{label} | x) \quad (1)
\]
If the subset of instances labeled by the oracle is representative of the interaction feature space, then many unknown interactions can be identified by the heuristic methods.

3. METHODS

3.1. Data

We used the dataset developed by Qi. et al. [2]. This dataset consists of 14,608 pairs of proteins that were known to interact. It also consists of 432,197 non-overlapping unlabeled instances. The protein-pairs (data instances) are represented by 27 features computed from biophysical characteristics of individual proteins. The 27 features correspond to: Gene Ontology (GO) cellular component, molecular function and biological process (3 features), co-occurrence in tissues (1 feature), gene co-expression (16 features), sequence similarity (1 feature), homology based (5 features) and domain interaction (1 feature). GO features measure the number of GO terms that are common in the annotations of the two proteins in the pair. As GO terms are categorized into three categories, namely the cellular component, molecular function and the biological process, the protein-pair features are computed separately for these three different similarity values. The tissue feature is a binary value indicating whether the two proteins have been expressed in the same tissue or not. This feature is added as it is observed that interacting proteins are likely to be expressed in the same tissue. 16 gene expression features were computed as correlations between gene expression values of the two genes in 16 different experiments. Sequence similarity is computed using BlastP sequence alignment E-value for the two proteins in the pair. Homologous proteins are obtained for each protein-pair in four different organisms namely yeast, fly, mouse and worm. This feature value is set to one, if the corresponding homologs are found to interact with each other in one or more of these organisms. Further details about the features in this dataset are described in the original source by Qi. et al. [2].

This dataset consisted of 14,608 interactions and 432,197 random pairs. From the random pairs, we created a subset of instances from this data such that every pair has more than 50% feature coverage. This was done so as to maintain a balance of feature coverage between interacting and random pairs, as the interacting pairs had a better feature coverage than the random pairs [10]. This subset is used for the development and the evaluation of the proposed methods. This subset has 180,800 protein-pairs in total. All the known protein interactions in the original set were included in this subset. 160,800 protein-pairs were selected randomly from this dataset for training and another 20,000 for testing. The test-data contained 5% interactions, which constitute 1,000 interacting protein-pairs, while the training data contained the remaining 13,608 interactions. A skewed dataset distribution is used to mimic the realistic scenario.

3.2. Base Classifier

Previously, Bayesian classifiers, logistic regression, support vector machines, decision trees and random forest have been proposed as supervised learning classifiers for PPI prediction [2, 16, 17]. It has been shown that random forest is best suited for this domain [2]. We used a Random forest containing 20 trees built by choosing from 8 different random features.

3.3. Oracle Simulation

We simulate the oracle behavior for PPI predictions using the set of known interactions that was downloaded from the human protein reference database (HPRD). The simulated oracle would assign the “interacting” label to a data-point if the protein-pair associated with the point is listed in HPRD. HPRD lists about 38,000 interactions pooled from various experimental sources. This list forms about 5% of all possible human PPIs [3]. Thus HPRD can be thought of as a reluctant one class oracle which gives the labels for only 5% of all the interacting class.

3.4. Accuracy Metrics

Precision, recall and F-score of positive class will be plotted as a function of the total cost in active learning. Precision is measured as the fraction of correctly predicted protein interactions among all the pairs predicted by the classifier to be interacting. Recall is the fraction of the interacting protein-pairs which the classifier is able to correctly identify as interacting pairs. F-score is the harmonic mean of precision and recall. F-score measures the accuracy of the method by combining both precision and recall values. Hence it can be used as a measure to compare the accuracy of the methods.

3.5. Algorithms

We propose active learning with a one class reluctant oracle, which attempts to learn the behavior of the oracle, and uses it to estimate the positive labels that are associated with unlabeled points by estimating \( P(\text{label} | x) \). We present three different ways to estimate the missing class for selected data points, which can be used in conjunction with any underlying active learning method. In the following methods, clusters are created using k-means clustering with Euclidean distance.

A. Estimate \( P(\text{label} | x) \) using number of interactions uncovered from each cluster (‘cluster-interactions’)
In this method, \( P(label \mid x) \) is estimated based on the number of labeled instances obtained from each cluster during the active learning iterations. The protein-pairs are clustered based on the existing feature-space, by considering all the features. If an oracle is able to provide the labels for many points in a cluster, it is most likely to do so for the other points too. That is, if the oracle has a 25% recall value, is able to give the labels for even 15% of all protein-pairs in a cluster, then in reality 60% (i.e. 15*1/0.25) other unlabeled points in the same cluster are more likely to be interacting points. Based on this principle, we propose a new metric to estimate labels of unlabeled points in a cluster, based on the size of the cluster, the recall value of the oracle, and the total number of positive instances that have been uncovered from the cluster till the current iteration. This value increases with increase in the number of labeled instances from every cluster. \( P(label \mid x) \) is thus estimated adaptively, during every active learning iteration as follows:

\[
P(label \mid x) = (1/Z) * P_k/L_k \quad \ldots (2)
\]

where, \( Z \) is a parameter whose value is between 0 and 1 and is chosen proportional to the recall of the oracle. As interacting class is a rare category in protein-protein interactions, if the oracle is highly reluctant, then setting value of \( Z \) to larger values would prove to be beneficial. \( k \) is the cluster to which point \( x \) belongs, \( P_k \) is the total number of labels obtained from the reluctant oracle, that is, the number of interactions that are uncovered from the cluster \( k \). \( L_k \) is the total number of data-points belonging to the cluster \( k \), whose labels are asked for by the system.

**B. Estimate \( P(label \mid x) \) using distance from known uncovered interactions (‘distance-interactions’)**

This method assigns \( P(label \mid x) \) based on the distance of the unlabeled data-point, from known uncovered interactions.

\[
P(label \mid x) = Z * (d_m - \|x_i - x\|) / d_m \quad \ldots (3)
\]

where \( x_i \) is the nearest labeled interaction from the data-point \( x \) that has been added to the training instances during the course of the active-learning iterations, \( \|x_i - x\| \) is the Euclidean distance between the points \( x_i \) and \( x \); and \( d_m \) stands for the maximum distance between any data-point \( x \) and its closest interaction instance \( x_i \).

Using this approach, the closer the points are to known interacting data-points, the more likely is the chance that they would be labeled as interactions. As distance from known interacting protein-pairs increases, this probability value reduces significantly. \( Z \) is a constant value between 0 and 1 which is assigned based on the recall value associated with the oracle.

**C. Estimate \( P(label \mid x) \) using \( \epsilon \)-neighborhood of uncovered interactions (‘\( \epsilon \)-neighborhood’)**

This method is based on the assumption that points closer to interacting data-points are more likely to be interacting in nature. We assign a label of “interacting” to all those unlabeled data-points shortlisted by any traditional active learning algorithm which falls in the \( \epsilon \)-neighborhood of known uncovered interactions. The \( \epsilon \)-neighborhood \( NN(x) \) of an instance consists of all those points located at most a distance of \( \epsilon \) from the data-point.

\[
NN(x) = \{ y \mid y \in U, \|x - y\| \leq \epsilon \} \ldots (4)
\]

Intuitively it could be thought of as a set of all data-points encompassed by a sphere with radius \( \epsilon \), drawn from the considered instance. For other points that do not fall in the \( \epsilon \)-neighborhood of known interactions, we assign them to be “non-interacting” in nature.

**Creation of a Weighted-Dataset**

We also experimented with creation of a weighted-dataset. Instead of assigning the “interacting” label with a probability of \( P(label \mid x) \) to the unlabeled data-point, and then adding it to the training set, we can create a weighted-dataset, by weighing unlabeled examples with a probability \( P(label \mid x) \). That is, each unlabeled example is considered to be “interacting” with a weight \( P(label \mid x) \) and “non-interacting” with a weight \( 1-P(label \mid x) \). Labeled data points are considered to be “known” interactions with a unit weight.

During the Active learning process, labels for select data-points are requested from the oracle. For all those points for which the label is obtained, a unit weight is assigned to each of them and they are added back to the training set. The unlabeled examples are duplicated. One copy is made as “interacting” with weight \( P(label \mid x) \) and the other is made as “non-interacting” with weight \( 1-P(label \mid x) \). Both these copies are added to the training set and a classifier is trained on the same. This entire process is repeated in every iteration.

We identify two separate methods namely, distance-interactions-weighted and distance-cluster-weighted which estimate \( P(label \mid x) \) using the distance-interactions and the distance-cluster based methods respectively, but create a weighted-dataset instead of directly estimating the missing class.

4. **RESULTS**

As a first step in evaluating the one class active learning methods, we determined the best possible values that could be associated with \( Z \) and \( \epsilon \) for all the three heuristics using a tenfold cross-validation technique on the training data. The values of \( Z \) and \( \epsilon \) were chosen in such a manner that those values contributed to the highest increase in F-score, while maintaining a high value of precision above a chosen threshold of 60%. The details of the cross-validations carried out to select \( Z \) and \( \epsilon \) (for each of the heuristic methods separately) are given in Supplementary File 1.
4.1. Evaluation on Test Data

After finding suitable values of $Z$ and $\varepsilon$ for each of the three heuristic methods separately, we evaluated the proposed one class heuristics on the held out data set and compared the performance against traditional active learning baseline and with random selection method. Uncertainty based active learning algorithm was used as the baseline, and 100 points were selected in every iteration. In the first iteration, 50 points are selected randomly and an initial classifier is built. Data is clustered into 50 clusters for the cluster-interactions method. Precision, recall and F-scores are computed for every iteration as shown in Figure 3.

It is observed that all of the active learners achieve the same F-score with only 1,500 instances, as the random selector does with 18,000 instances. Also the one class active learning methods have a higher value of F-score as compared to the baseline of uncertainty based active learning. For the parameters selected through cross-validation for each method, cluster-interactions method has the highest value of F-score, followed by the distance-interactions method and the $\varepsilon$-neighborhood method. The $\varepsilon$-neighborhood method has the same F-score as that of the baseline active learner for about 10,000 instances, after which the F-score associated with the baseline approach starts decreasing slightly. This decrease can be attributed to a biased selection of positive instances during the initial iterations of the active learning process. In the later iterations, more number of random unlabeled instances are picked for which the oracle is not able to provide the label and are added to the training set as negative data, which results in a small decrease in the F-score. However, the one class methods do not consider these points to be negative in nature. An appropriate label is assigned to such points using the heuristics proposed above. Some of the unlabeled data get labeled as positive instances due to which the F-score remains the same for these one class methods.

The challenge is to significantly improve upon recall of the entire system without compromising on its precision. The one class methods have a slightly lower value of precision but have much better values of recall, which can be observed through their increased F-scores. In the domain of PPI prediction, where the positive class is a very rare category, an improvement in recall is more difficult to attain than improvement in precision (because a slight inaccuracy in classifier could potentially misclassify an order of magnitude more of negative instances into positive class, thereby dropping precision significantly). Note that the reported precision and recall are computed for the positive class as is typical in this domain.

Figure 4 shows the results associated with the weighted-dataset methods. These methods are compared against the traditional active learning baseline as well as their corresponding counterparts that estimate potential labels before adding the data-point to the training set. Weighted-dataset methods are similar to the other one class active learning methods discussed previously. The interactions-distance weighted method has the same precision, recall and F-score trends as its one class active learning counterparts. The interactions-cluster weighted method has a very poor value of precision, and a very high value of recall as compared to its one class counterpart for the initial few iterations, and bounces back to, thus having a similar trend in the F-scores.

5. DISCUSSION

In the domain of protein-protein interaction prediction, or any computational biology task in general, the few instances of labeled data currently known are often insufficient to confidently characterize remaining unlabeled data. Conversely, data characterization using wet-lab experiments is expensive in terms of expert
manpower, time, and resources. When choosing a specific data instance (e.g. molecule or protein-pair) to be studied by wet-lab experiments, its redundancy with previously annotated (labeled) data is rarely taken into account. The choice of which molecule is to be studied is mostly based on the expertise of the lab proposing to study it and on availability of reagents that make it possible to study with the said experimental methods. However, the trend today is moving towards high-throughput techniques (HT) such as yeast 2-hybrid technology for the study of PPIs. Even with HT, it is not feasible to study every molecule or protein-pair. Computational methods are therefore required to predict the annotations. Given the availability of computational methods and HT techniques, it is desirable to have active learning algorithms that guide the selection of some data which when labeled by HT methods improves the accuracy and confidence of labeling the remaining data with computational methods.

We proposed one class active learning heuristics in this paper which deal with a one class reluctant oracle for the domain of protein-protein interaction prediction. They estimate the potential label associated with shortlisted data-points during the active learning process for which the oracle does not give the label. These one class methods have a higher recall and F-score as compared to the traditional active learning counterparts. There is a slight loss in precision, but this value of precision is in itself an underestimate as there is no gold standard dataset for evaluating protein interactions, and false positives in the test set may actually turn out to be true undiscovered interactions.

Although similar to the problem of learning from positive and unlabeled data as in other domains (e.g. document classification [18-20]), PPI discovery is a rare-category problem with large instance space and very small set of highly related features. Traditional approaches to dealing with positive and unlabeled data have already been applied to PPI prediction, and the goal of this work is to discover more interactions beyond those that are already predicted by treating the problem as positive and unlabeled data. Previous methods have
treated learning from positive and unlabeled data as standard. By comparing the methods proposed here with regular active learning and random selection of training data, we show that estimating the label for some of the training data provides superior results.

While proposing these heuristics, we made certain assumptions associated with the underlying domain. We assume that the underlying PPI feature space can characterize oracle behavior; and that the oracle should provide labels that make up a randomly selected subset of the entire interaction space to uncover all potential undiscovered interactions. Although these assumptions may not hold in certain circumstances, given that oracle characterization for PPI prediction is a non-trivial problem, we believe that the one class heuristics proposed in this paper are the stepping stone for solving this problem.

These methods help to incorporate semi-supervised learning approaches to extend active learning to more realistic scenarios by estimating the labels of points for which the oracle does not give the label. These approaches improve the recall of the prediction system.

By characterizing the oracle behavior, we are proposing to use it to exploit the strengths of any experimental method and present an approach for rapid development of the interactome. This would provide a sound basis for rapid discovery of interactomes under budget constraints. Collaborations are underway towards labeling selected instances with yeast 2-hybrid methods.

6. ACKNOWLEDGEMENTS

Funding: MG and SA’s work has been funded by the Biobehavioral Research Awards for Innovative New Scientists (BRAINS) grant R01-MH094564 awarded to MG by the National Institute of Mental Health of National Institutes of Health (NIMH/NIH) of USA.

REFERENCES

15. Elkan, C. and K. Noto. Learning classifiers from only positive and unlabeled data. in Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008. Las Vegas, Nevada, USA: ACM.