Exploiting Unlabeled Data for Improving Accuracy of Predictive Data Mining

Kang Peng, Slobodan Vucetic, Bo Han, Hongbo Xie, Zoran Obradovic
Center for Information Science and Technology, Temple University, Philadelphia, PA 19122, USA
{kangpeng, vucetic, hanbo, hongbox, zoran}@ist.temple.edu

Abstract

Predictive data mining typically relies on labeled data without exploiting a much larger amount of available unlabeled data. The goal of this paper is to show that using unlabeled data can be beneficial in a range of important prediction problems and therefore should be an integral part of the learning process. Given an unlabeled dataset representative of the underlying distribution and a K-class labeled sample that might be biased, our approach is to learn K contrast classifiers each trained to discriminate a certain class of labeled data from the unlabeled population. We illustrate that contrast classifiers can be useful in one-class classification, outlier detection, density estimation, and learning from biased data. The advantages of the proposed approach are demonstrated by an extensive evaluation on synthetic data followed by real-life bioinformatics applications for (1) ranking PubMed articles by their relevance to protein disorder and (2) cost-effective enlargement of a disordered protein database.

1. Introduction

A common assumption in supervised learning is that labeled data conforms to the same distribution as the data to which the predictor will be applied. However, due to various reasons such as sampling bias or prohibitive labeling costs, labeled datasets are often small and/or biased, which makes them unrepresentative of the underlying distribution. A predictor learned from such data may not generalize well on out-of-sample examples. On the other hand, it is typically easier to collect large amounts of unlabeled data at a significantly lower cost. Moreover, unlabeled samples are less likely to be biased and could therefore often be considered as representatives of the underlying data distribution. This property of unlabeled data makes it an attractive tool for improving accuracy of predictive data mining [17].

As a natural approach to handling unlabeled data, the Expectation Maximization (EM) algorithm [6] can be used to iteratively estimate the model parameters and assign soft labels to unlabeled examples by treating the unknown labels as missing data and assuming generative model such as mixture of Gaussians. EM has been widely used in areas such as text document classification [10], image retrieval [22] and multispectral data classification [18]. Co-training [2] provides another popular strategy for incorporating unlabeled data if the data can be described in two different sufficient views, or sets of attributes. The transduction approach [20] assigns labels to a given set of unlabeled data by maximizing the classification margins on both labeled and unlabeled data. However, it is often observed that these techniques could also degrade performance due to violated model assumption or convergence to local maxima [17].

Another line of research deals with learning problems characterized by extreme bias in labeled data. In one-class classification problems [19] only labeled examples from a single class are available and the goal is to predict out-of-sample examples either as belonging to the class or as outliers. A possible approach to address this problem is to apply kernel density estimation (KDE) for learning the probability density of labeled data. Another approach is the support vector data description (SVDD) method which learns from the positive examples and artificially generated outliers [19] to separate the positive class from the rest. However, these approaches are ignoring unlabeled data that might be readily available. Partially supervised classification [9] provides a notable solution to one-class classification problems that utilizes unlabeled data; it initially assumes that all unlabeled examples come from the negative class, and then applies the EM algorithm to refine the assumption.

In this study, we propose a novel approach for utilizing unlabeled data in data mining. It is based on constructing a *contrast classifier* that discriminates between labeled and unlabeled examples. The name contrast classifier comes from the meaning of its output - it represents a measure of difference, or contrast, in density of a given data point between labeled and unlabeled data. Given this property, it is apparent that contrast classifiers could be used in a wide range of important data mining applications such as outlier detection, one-class classification, density estimation, and learning from biased data. In Section 2, we provide a description of contrast classifiers and a range of their applications. In Section 3 we compare our approach to alternative methods on a challenging 3-class synthetic dataset. In

Section 4, we illustrate the usefulness of the contrast classifiers on two real-life bioinformatics problems of (1) re-ranking of PubMed articles based on their relevance to protein disorder and (2) cost-effective enlargement of a dataset of disordered proteins.

2. Methodology

2.1. Contrast classifiers

By $g(\mathbf{x})$ we denote the probability density function (pdf) of unlabeled data $\mathbf{U} = \{\mathbf{x}_i, i = 1, \ldots, \mathrm{N_U}\}$, and we assume that it corresponds to the underlying distribution. In a K-class classification problem, $g(\mathbf{x})$ can be represented as a mixture of K class-conditional pdfs, or $g(\mathbf{x}) = \Sigma_j p_j g_j(\mathbf{x})$, where p_j is prior probability of class j and $g_j(\mathbf{x})$ is class-conditional pdf of data from class j. In a number of applications, the available unlabeled dataset $\mathbf{U}$ is large and could be used to derive a quite accurate estimate of $g(\mathbf{x})$. By $h(\mathbf{x})$ we denote the probability density function of labeled data $\mathbf{L} = \{(\mathbf{x}_i, \mathrm{c}_i), i = 1, \ldots, \mathrm{N_L}, \mathrm{c}_i \in \{1, 2, \ldots, K\}\}$. It can also be represented as a mixture of class-conditional pdfs, or $h(\mathbf{x}) = \Sigma_j q_j h_j(\mathbf{x})$, where q_j is prior probability of class j and $h_j(\mathbf{x})$ is class-conditional pdf of data from class j. Since $\mathbf{L}$ could be obtained through biased sampling, $h_j(\mathbf{x})$ may not equal $g_j(\mathbf{x})$ and thus $h(\mathbf{x})$ may not equal $g(\mathbf{x})$.

We define a *contrast classifier* $cc(\mathbf{x})$ as a classifier trained to discriminate between labeled data (class 0) and unlabeled data (class 1). Given an input $\mathbf{x}$, the optimal contrast classifier able to approximate posterior class probability would output

$$cc(\mathbf{x}) = \frac{r \cdot g(\mathbf{x})}{(1-r) \cdot h(\mathbf{x}) + r \cdot g(\mathbf{x})}\,, \qquad (1)$$

where r is the fraction of unlabeled data in the training set. In the following subsection we will discuss the proper choice of r in practical applications. It is evident that if unlabeled and labeled data are drawn from the identical distribution, i.e., $g(\mathbf{x}) = h(\mathbf{x})$, the optimal $cc(\mathbf{x})$ would be a constant equal to r.

Assuming that the pdf of unlabeled data $g(\mathbf{x})$ is known, the optimal $cc(\mathbf{x})$ can be used to estimate $h(\mathbf{x})$ as

$$h(\mathbf{x}) = \frac{1 - cc(\mathbf{x})}{cc(\mathbf{x})} \cdot \frac{r}{1-r} \cdot g(\mathbf{x})\,. \qquad (2)$$

To measure the difference in density of labeled and unlabeled data, we define the *contrast* as ratio $h(\mathbf{x})/g(\mathbf{x})$ and from equation (2) we have

$$contrast(\mathbf{x}) = h(\mathbf{x}) / g(\mathbf{x}) = \frac{1 - cc(\mathbf{x})}{cc(\mathbf{x})} \cdot \frac{r}{1-r}\,. \qquad (3)$$

As we will illustrate later, the contrast measure can be extremely useful in a number of applications such as one-class classification, outlier detection and learning from biased data. Since *contrast*($\mathbf{x}$) is a monotonically decreasing function of $cc(\mathbf{x})$, contrast classifiers could be used directly to rank examples by their contrast measure.

If labeled data consist of K classes, K *class-specific* contrast classifiers $cc_j(\mathbf{x})$, j=1, 2, …, K, could be constructed where $cc_j(\mathbf{x})$ is trained to discriminate between unlabeled data and class j of labeled data. The class-specific contrast classifiers can then be used to construct the *maximum a posteriori* (MAP) classifier. Using (2), posterior probability $p(c = j \,|\, \mathbf{x})$ of class j can be expressed as

$$p(c = j \,|\, \mathbf{x}) = \frac{h_j(\mathbf{x}) \cdot q_j}{\sum_i h_i(\mathbf{x}) \cdot q_i} = \frac{q_j \cdot (1 - cc_j(\mathbf{x})) / cc_j(\mathbf{x})}{\sum_i q_i \cdot (1 - cc_i(\mathbf{x})) / cc_i(\mathbf{x})}\,, \qquad (4)$$

where, for simplicity, r was set to 0.5 in all $cc_j(\mathbf{x})$, j = 1, …, K. Often, we have some knowledge about prior class probabilities in unlabeled data and about the misclassification costs, which should be used to select more appropriate priors in (4) instead of q_j. Based on (4), the decision rule using contrast classifiers as a MAP classifier can be expressed as

$$\hat{c} = \arg\max_j \frac{(1 - cc_j(\mathbf{x}))}{cc_j(\mathbf{x})} \cdot q_j\,. \qquad (5)$$

An important result is that knowledge of g($\mathbf{x}$) is not needed to construct the MAP classifier from class-specific contrast classifiers. Therefore, as seen from (5), if labeled data are an unbiased sample, contrast classifiers can be directly used to provide an optimal solution for multi-class problems.

2.2. Construction of contrast classifiers

In practice, unlabeled dataset could be much larger than labeled dataset. Learning on such imbalanced data would result in a low-quality contrast classifier, while learning time could be prohibitively long. Moreover, the quantity $(1\text{-}cc(\mathbf{x}))/cc(\mathbf{x}) \cdot r/(1\text{-}r)$, which provides a measure of contrast in labeled data, is numerically less stable for imbalanced training data with large r. Learning on imbalanced data has received a lot of attention in the data mining community, and most successful strategies are based on balanced training data [8]. In our approach, we train an *ensemble* of classifiers on *balanced* training sets consisting of equal number of labeled and unlabeled examples randomly sampled from the available labeled and unlabeled data. Similar to bagging [5] we construct a contrast classifier by aggregating the predictions of these classifiers through averaging. With the proposed method we are effectively using information available from the large unlabeled data, thus allowing construction of more accurate contrast classifiers.

Classification algorithms able to approximate posterior class probability are suitable choices for contrast classifiers. These algorithms include logistic regression, feed-forward neural networks with sigmoid activation functions [15], or even decision trees [14]. Additionally, there are several types of binary classifiers (e.g. support vector machines; naive Bayes) producing output scores that can be interpreted as prediction strength. It has been shown that these scores could be successfully calibrated to posterior class probability in case of support vector machine (SVM) [12] and naive Bayes [24] classifiers. The results indicate that the output from such classifiers can be interpreted as a monotonically increasing function of the posterior class probabilities. It follows that $h(\mathbf{x})/g(\mathbf{x})$ can be represented as $\mathrm{F}(cc(\mathbf{x}))$, where $\mathrm{F}(\cdot)$ is a monotonically decreasing function. In section 4 we will illustrate that this property allows successful use of classifiers such as SVM as contrast classifiers in some important applications.

While contrast classifiers could be used to produce a MAP classifier, it is evident that $(1\text{-}cc(\mathbf{x}))/cc(\mathbf{x})$ term in (4) could cause prediction instability for inputs $\mathbf{x}$ with small $cc(\mathbf{x})$. In practice, this should not be an issue since unlabeled data can be considered to be a mixture of distributions with one or more mixture components corresponding to the unlabeled data. Therefore, contrast classifiers are not likely to produce values near zero. Exceptions could be extreme scenarios where labeled data are outliers obtained by highly biased sampling and covering a highly limited portion of a feature space. The property of unlabeled data that allows construction of well-behaved contrast classifiers in most realistic cases is essential for the success of the proposed methodology.

2.3. Contrast classifiers for density estimation

As seen from equation (2), contrast classifiers could be used for class-conditional density estimation. Since unlabeled dataset could be very large, relatively accurate estimation of $g(\mathbf{x})$ should be feasible using some of the standard nonparametric methodologies such as mixture modeling or kernel density estimation. Therefore, if a contrast classifier able to approximate posterior class probability is used, it should be possible to obtain a fairly accurate estimate of labeled data pdf from (2). However, in this paper we will not pursue this direction any further.

2.4. Contrast classifiers for one-class classification and outlier detection

In one-class classification, labeled dataset contains examples from only one class, called *positive* class. The goal is to build a model able to recognize whether a new example is from positive class. Complementary, the task can be detection of outlying examples that are distributionally underrepresented in labeled data. Therefore, *contrast*(**x**) from equation (3) is very suitable for both tasks: given an appropriate threshold, all examples with *contrast*(**x**) above (below) the threshold can be classified as positive (outliers). Since *contrast*(**x**) is a monotonically decreasing function of $cc(\mathbf{x})$, a threshold can be applied directly on contrast classifier outputs.

A choice of an appropriate threshold for one-class classification could be difficult since the output range of contrast classifiers depends on the dataset. To alleviate this problem, in our approach we introduce the threshold θ^p such that condition $cc(\mathbf{x})>\theta^p$ is satisfied for a user-specified p% of labeled examples. Therefore, p represents an upper bound on the false negative rate (percent of rejected positive examples), and a user should select it in order to achieve the optimal trade-off between false positive and false negative rates in one-class classification.

2.5. Contrast classifiers for learning from biased data and generalized outlier detection

As seen from equation (4), contrast classifiers could in theory be used to substitute standard multi-class classification algorithms. Therefore, in scenarios where labeled data is an unbiased sample with $h(\mathbf{x}) = g(\mathbf{x})$ both approaches should achieve the similar classification accuracy. However, in the more general setup where labeled data is a biased sample with $h(\mathbf{x}) \neq g(\mathbf{x})$, the benefits of contrast classifiers become apparent. They follow from the ability of contrast classifiers to detect examples underrepresented in labeled data while achieving near-optimal classification on the others.

For a given example $\mathbf{x}$, let $cc_j(\mathbf{x})$, $j = 1, 2, \ldots, K$, denote the outputs of K class-specific contrast classifiers. If all K outputs are large, $\mathbf{x}$ is likely to be an outlier or an example underrepresented in the labeled data. In such a case, the best policy could be not to provide classification; this would result in an increased overall accuracy at the cost of somewhat decreased coverage. Similar to the use of contrast classifiers in one-class classification, in our approach a user-specified constant p is used to determine K thresholds θ_j^p, $j = 1, 2, \ldots, K$, such that p% of positive training examples satisfy $cc_j(\mathbf{x}) > \theta_j^p$, for each $j = 1, 2, \ldots, K$. Classification is not provided for examples with $cc_j(\mathbf{x}) > \theta_j^p$ for all $j = 1, 2, \ldots, K$. Otherwise, equation (4) is used for classification.

It is evident that the procedure for classification on biased data includes detection of outliers that we denote as *generalized outlier detection* since it detects examples underrepresented in each of the K classes available in labeled data.

3. Experiments on waveform data

In this section we use the well-known waveform dataset [4] to illustrate the effectiveness of our approach on one-class classification in the presence of unlabeled data, as well as on multi-class classification on unbiased and biased data. In this 3-class dataset, there are 21 attributes defined as a linear combination of two out of 3 basic waveforms with randomly generated coefficients. Its noisy version includes 20 additional irrelevant attributes with Gaussian distribution. Learning on the waveform dataset is generally considered a difficult task with reported accuracy of 86.8% using a Bayes optimal classifier.

3.1. One-class classification

We first compared contrast classifier (CC) with two alternatives for one-class classification: kernel density estimation (KDE) and support vector data description (SVDD) [19]. While both KDE and SVDD learn exclusively from labeled data to directly or indirectly estimate $h(\mathbf{x})$, the contrast classifier utilizes unlabeled data to estimate $h(\mathbf{x})/g(\mathbf{x})$.

Kernel density estimators directly estimate $h(\mathbf{x})$ from labeled data as

$$h(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} G(\mathbf{w};\mathbf{x}-\mathbf{x}_i) \qquad (6)$$

where G is the *Gaussian* kernel with bandwidth **w** and n is the number of labeled examples. If a new example has $h(\mathbf{x})$ below a certain threshold it is considered an outlier. The threshold can be determined as described in Section 2.4. The optimal bandwidth w is the value that maximizes the data likelihood.

Instead of estimating the $h(\mathbf{x})$ directly, the SVDD method uses artificially generated outliers [19] along with the positive examples to construct a hyperspherical decision boundary of minimal possible volume that separates the positive class examples from others. To construct more flexible decision boundaries, kernel functions are introduced to map the data into a higher dimensional space. For SVDD experiments we used the data description toolbox (*dd_tools*) from *http://ida.first.gmd.de/~davidt/*.

The contrast classifier was implemented as an ensemble of feed-forward neural networks. Each network had 10 hidden and 1 output sigmoid neurons. The number of component neural networks was determined empirically. To train a single network, a balanced dataset was formed from N examples taken randomly with replacement from the labeled data and another N from the unlabeled data. As discussed in Section 2.4, we used contrast classifier output $cc(\mathbf{x})$ instead of $contrast(\mathbf{x})$ for classification.

Two sets of waveform data were generated: N (<< 150,000) labeled examples from class 1 and 150,000 unlabeled examples, 50,000 from each of the 3 classes. To examine the effect of labeled data size and irrelevant attributes, experiments were performed under four scenarios: 1) N = 200, 2) N = 200 noisy, 3) N = 2000, 4) N = 2000 noisy. Here "noisy" refers to the noisy version of waveform data with 20 irrelevant attributes. The accuracy of one-class classification was measured as the true positive rate when the false positive rate was 20%. For CC and KDE, the desired false positive rate was obtained by selecting appropriate thresholds. For SVDD, it was obtained from the ROC curve generated by the *dd_tools* software.

In Figure 1 we show that the accuracy of contrast classifiers improved with the number of component neural networks but then saturated at around 20. Therefore, in the remaining experiments in Section 3 we used an ensemble of 20 neural networks as the contrast classifier. Table 1 compares the accuracies of the three methods. In all four scenarios, contrast classifier was superior to the other two methods showing that unlabeled data could greatly improve the accuracy of one-class classification. It is worth noting that the performance of the KDE method degraded as irrelevant attributes were introduced, while CC and SVDD appeared to be robust to noise.

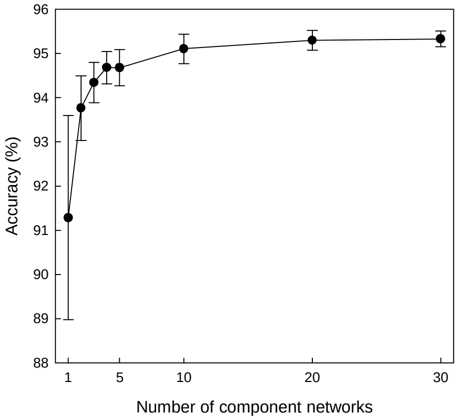


Figure 1. The effect of the number of component neural networks on accuracy for N = 2000 noisy scenario

3.2. Classification with unbiased data

As shown in Section 2.1, if labeled data is unbiased, class-specific contrast classifiers could be used to construct a MAP classifier. In this section, we compare it with a standard 3-class neural network classifier, which is an ensemble of 20 3-output neural networks each trained on a bootstrap replicated sample of the labeled data. The 3 class-specific contrast classifiers were trained with a balanced set consisting of labeled examples from a given class and unlabeled examples. For this experiment, a total of 1500 labeled examples and 150,000 unlabeled examples were generated with different class proportions.

In Table 2 we report their accuracies obtained in experiments with 3 different class proportions. The overall accuracy was calculated as average of individual class accuracies weighted by class proportions. It is evident that the contrast classifier approach achieved accuracy comparable to the best multi-class classifiers when labeled data is unbiased. As will be seen in the next subsection, the true strength of contrast classifiers becomes apparent when labeled data is biased.

Table 1. Comparison of three methods in one-class classification

Dataset	Method	Accuracy (%)
N = 200	CC	94.9
	KDE	63.8
	SVDD	67.7
N = 200 noisy	CC	92.7
	KDE	54.3
	SVDD	68.6
N = 2000	CC	96.1
	KDE	65.2
	SVDD	68.1
N = 2000 noisy	CC	95.1
	KDE	56.9
	SVDD	65.3

CC - contrast classifier, KDE - kernel density estimation, SVDD - support vector data description

Table 2. Comparison of a MAP based on class-specific contrast classifier and a standard 3-class neural network in classification with unbiased labeled data

Class Proportion	Method	Class 1 (%)	Class 2 (%)	Class 3 (%)	Overall (%)
1:1:1	CC	81.9	88.1	88.4	86.1
	NN	80.9	87.1	89.4	85.8
1:4.5:4.5	CC	54.9	94.4	94.2	90.4
	NN	65.2	91.1	94.7	90.1
1:1:8	CC	62.9	75.0	99.5	93.4
	NN	69.7	78.6	98.5	93.6

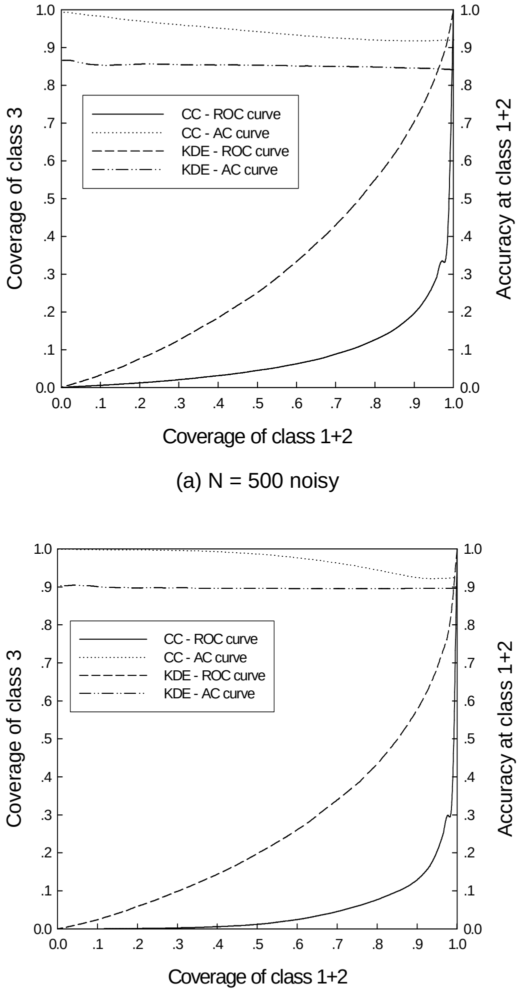
CC - the MAP classifier based on 3 class-specific contrast classifiers
NN - an ensemble of 20 neural network with 3 outputs

3.3. Classification with biased data

We consider a biased data scenario where examples from class 3 are completely missing from the labeled data. In such a case, the desired classifier should have high classification accuracy on examples from classes 1 and 2, while it should be able to recognize class 3 examples as underrepresented in labeled data and thus refuse to predict on them. We examined the performances of contrast classifier (CC) and kernel density estimation (KDE) approaches on this challenging problem.

For the CC approach, two contrast classifiers specific for classes 1 and 2 were constructed and combined to detect underrepresented examples as described in Section 2.5. For the KDE approach, the class-conditional densities $h(\mathbf{x}|c{=}1)$ and $h(\mathbf{x}|c{=}2)$ were estimated from the labeled data. If $h(\mathbf{x}|c{=}j) < \theta_j^p$ for both j=1, 2, **x** was characterized as an outlier and classification was not provided. The thresholds θ_j^p were determined such that p% of class j examples satisfy $h(\mathbf{x}|c{=}j) < \theta_j^p$. Otherwise, classification was provided using the Bayes rule: if $h(\mathbf{x}|c{=}1) > h(\mathbf{x}|c{=}2)$, **x** was labeled with class 1 and vice versa.

We generated labeled data with N examples from classes 1 and 2. The unlabeled data consisted of 50,000 examples for each of the 3 classes. Two experiments were performed: (a) N = 500 with 20 noisy attributes, (b) N = 5000 without noisy attributes. For a range of choices of parameter p with both CC and KDE approaches we measured (1) classification accuracy of classes 1 and 2, (2) prediction coverage of classes 1 and 2, and (3) prediction coverage of class 3. An ideal predictor should have a 100% accuracy and 100% coverage of classes 1 and 2, but 0% coverage of class 3.



(a) N = 500 noisy

(b) N = 5000

Figure 2. ROC and AC curves for classification with biased data

In Figure 2 we report the performance of both approaches as: (1) ROC curve - class 3 coverage vs. class 1+2 coverage (2) AC curve - accuracy vs. class 1+2 coverage. Clearly, CC achieves both better accuracy and lower class 3 coverage than KDE for a whole range of class 1+2 coverage. As in Figure 2(b), while retaining 95% coverage on classes 1+2, CC approach reduced class 3 coverage to about 20% vs. 70% by KDE, thus was more effective in detecting outliers. An interesting result is that slight increase in accuracy was achieved with decrease in class 1+2 coverage in both scenarios with both models. Consistent with results reported in Section 3.1, contrast classifiers proved to be very robust to noisy attributes and small labeled data size. These results show that unlabeled data can be extremely useful in classification of biased labeled samples and should be an integral part of learning process whenever available.

4. Bioinformatics application: analysis of protein disorder

Disordered proteins are characterized by long regions of amino acids that do not have a stable three-dimensional conformation under normal physiological conditions. Recent results indicate that, despite the traditional view, disordered proteins are common in nature and are responsible for a spectrum of important biological functions [7]. However, due to the historical overlooking of this property, the knowledge about protein disorder is scattered across literature and described with non-unified terminology. Important data mining challenges include allowing cost-effective extraction of knowledge about protein disorder from literature, as well as assisting in better understanding available information about protein disorder. In this section we illustrate that contrast classifiers are appropriate tools for addressing these challenges.

4.1. Ranking of PubMed articles

In a search for biological papers describing properties of uncharacterized disordered proteins, we started from a set of 178 articles describing properties of 90 known disordered proteins collected through intensive literature search by several experts [7]. By querying PubMed (*http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed*), an open access web-based archive of biomedical literature, with names of the 90 disordered proteins, we found that 67 of these proteins had more than 100 PubMed citations, 35 had more than 1,000 citations, and 12 had more than 10,000 citations. Out of the 178 articles only 13% (28%) were returned as the top 5 (10) PubMed retrievals. Our goal was to improve this fraction by ranking the abstracts retrieved by PubMed based on their relevance to properties of disordered proteins extracted automatically from the relevant articles.

In this one-class classification scenario, the labeled (positive) set P contained 166 abstracts stored at PubMed, while the unlabeled set contained 18,499 abstracts from PubMed obtained by querying with the 90 disordered protein names. After removing infrequent and stop words, remaining words were preprocessed into terms by eliminating suffixes using Porter stemmer [13]. Frequencies of terms were computed over C and P sets and terms were ranked based on the difference in their frequency in C and P. The most discriminative K terms, called keywords, were used to represent each PubMed abstract as a vector of TF-IDF weights [16] calculated as a ratio between the term frequency and the inverse document frequency. Following the approach outlined in Section 2.2, we then trained the contrast classifier as a linear SVM to rank the unlabeled abstracts based on its output.

Using the top 200 keywords (K = 200), significant improvement was achieved in ranking as compared to PubMed default output: the fraction of citations ranked in the top 5 (top 10) was increased from 13% (28%) to 50% (71%). These results suggest that labor involved in finding relevant literature can be reduced many times through the proposed re-rankings by contrast. We note that while the illustrated use of unlabeled data in text mining is not novel, the value of our work is in describing this approach through the statistically appealing framework of contrast classifiers.

4.2. Contrast classifiers for study of protein disorder

Here the problem was to discover and understand proteins that are underrepresented in a labeled database of known ordered and disordered proteins. By using contrast classifiers we showed that the outlying proteins are numerous and have specific properties that may provide a novel insight into structural and functional properties of proteins.

4.2.1. Contrast classifiers for ordered and disordered proteins prediction. The labeled dataset we used consisted of 152 proteins containing disordered regions longer than 30 consecutive residues and 290 completely ordered proteins [21]. Every pair of the labeled sequences had less than 30% sequence identity. The unlabeled data was constructed from the October 2001 release 40 of *SWISS-PROT* database [3] containing 101,602 proteins. The ProtoMap database [23] was used to group these proteins into 17,676 clusters based on their sequence similarities [21]. One representative protein was then selected from each cluster, resulting in an unlabeled dataset of 17,676 proteins.

In our previous work it was found that order/disorder properties of a given sequence position could be predicted fairly accurately based on sequence properties within a symmetric input window centered on that position. Our currently best disorder predictor VL3 [11], an ensemble of 10 neural networks, uses 20 window-based attributes including 18 relative frequencies of 18 out of the 20 amino acids within an input window of length 41, the flexibility index averaged over the window, and the K2-entropy. Its overall accuracy is 83.9%, with 76.3%/91.4% accuracy on disorder/order class.

In a more recent study [21] two class-specific autoassociator neural networks were constructed to detect underrepresented proteins. The two resulting models were effective in discovering important classes of under-represented proteins. However, the overall accuracy of a disorder predictor based on the two models was only 69.8%, more than 10% worse than that of VL3, indicating that more accurate outlier detection is possible.

In this study, we built two class-specific contrast classifiers $cc_{disorder}$ and cc_{order} as ensembles of 50 neural networks using the same attributes as VL3. A MAP disordered predictor was then constructed according to equation (5) where both priors were set to 0.5. Its overall accuracy was 84.0%, with 75.6%/92.3% on disorder/order class, which were practically identical to those of VL3. This result suggests the effectiveness of contrast classifiers in the selection and analysis of underrepresented proteins.

4.2.2. Application of contrast classifiers to selection and analysis of underrepresented proteins. We first filtered the 17,676 unlabeled proteins by applying one round of *blastp* algorithm [1] with E-value threshold 1 to remove proteins similar to the labeled proteins. Then we retained only those with lengths between 200 and 500 amino acids, which resulted in the *SWISS* set with 6,964 proteins.

After applying the two contrast classifiers on a protein of length L, two L-dimensional vectors of position-by-position predictions are obtained. To allow detection of proteins that are overall the most different from the labeled ordered and disordered proteins, we summarized each protein with cc_avg_{order} and $cc\_avg_{disorder}$, representing the average predictions of the contrast classifiers. Similar to the approach described in Section 2.5, we determined thresholds $\theta_{order}{}^{p}$ and $\theta_{disorder}{}^{p}$, such that p% of *SWISS* proteins satisfy $cc\_avg_{order} > \theta_{order}{}^{p}$ and $cc\_avg_{disorder} > \theta_{disorder}{}^{p}$, respectively. In Figure 3 we show the proportions of selected outliers from *SWISS* set and the labeled proteins for different p. As could be seen, the proportion of outliers in *SWISS* set is significantly higher than the labeled proteins for a whole range of choices for p. This shows that a significant portion of proteins from *SWISS-PROT* have properties different from the known ordered and disordered proteins.

Using p = 50 we selected 1,259 outliers from *SWISS* proteins and denoted this set as *OutAvg*. To properly evaluate *OutAvg* proteins, we constructed two additional datasets: *OrdHom* with 539 *SWISS-PROT* homologues of 290 ordered proteins, and *DisHom* with 356 *SWISS-PROT* homologues of 152 disordered proteins. Similar to our previous approach [21], for each dataset we calculated the frequencies of 840 keywords listed in *SWISS-PROT* that summarize the structural and functional properties of a given protein. Table 3 shows a summary for the 3 most interesting keywords selected according to their frequencies, which correspond to a large family of membrane proteins known to have specific structural and functional characteristics. It can be seen that they are highly underrepresented among our labeled ordered and disordered proteins as compared to *SWISS*, while they are very common in the identified set of outliers *OutAvg*. Thus, it is likely that most of the bias in our labeled data comes from membrane proteins. It is the matter of further research to determine the full significance of these and other detected underrepresented functional groups and their impact on our study of protein disorder.

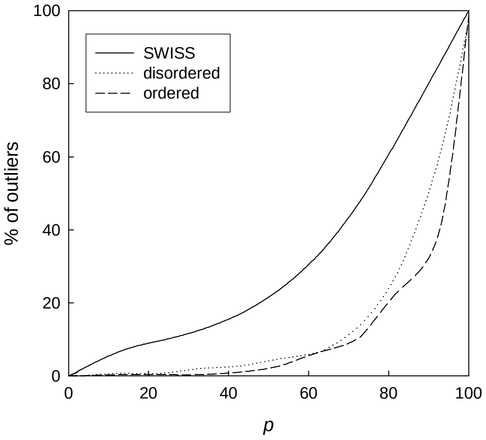


Figure 3. The proportion of selected outliers from *SWISS* set, disordered and ordered proteins for different *p*.

Table 3. Comparison of frequencies of the 3 most interesting keywords associated with proteins in 4 datasets.

Keyword	*SWISS*	*OrdHom*	*DisHom*	*OutAvg*
Inner Membrane	2.1	2.2	2.1	6.5
Membrane	21.1	13.4	13.2	57.6
Transmembrane	17.7	9.3	8.9	55.7

5. Conclusions

We proposed a framework for exploiting large amount of available unlabeled data in order to improve accuracies of various predictive data mining tasks such as one-class classification, outlier detection, and learning with biased data.

As the crucial part of our approach, the contrast classifiers are trained to characterize the contrast or difference between the possibly biased labeled data and unlabeled data. Performance of contrast classifiers was similar to standard classifiers when labeled sample is unbiased. However, the true strength of contrast classifier comes from its ability to effectively detect outlying examples with statistical properties contrasting those of labeled data. While the extensive experiments on synthetic data provided a useful characterization of the proposed framework compared to a range of standard alternatives, the two successful applications in biology domain showed that contrast classifiers could be very useful in solving important practical problems.

The conclusion is that unlabeled data, if available in large amount should be considered as an integral part of data mining process and, therefore, should not be ignored. The results indicate that the appropriate use of unlabeled data could be greatly beneficial to improvement of predictive data mining quality.

Acknowledgements

This study was supported in part by NSF grant IIS-0219736 to Z. Obradovic and S. Vucetic and NSF grant CSE-IIS-0196237 to Z. Obradovic and A.K. Dunker.

References

[1] S.F. Altschul, T.L. Madden, A.A. Schiffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”, *Nucleic Acids Res.*, 1997, vol. 25, pp. 3389-3402.

[2] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training”, In *Proc. of COLT’98*, 1998, pp. 92-100.

[3] B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout and M. Schneider, “The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003”, *Nucleic Acids Res.*, 2003, vol. 31, pp. 365-370.

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*, Wadsworth Inc., 1984, pp. 43-49.

[5] L. Breiman, “Bias, variance, and arcing classifiers”, Technical Report 460, UC-Berkeley, 1996.

[6] A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, *J. Roy. Statistical Society (B)*, 1977, vol. 39, pp. 1-38.

[7] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva and Z. Obradovic, “Intrinsic disorder and protein function”, *Biochemistry*, 2002, vol. 41(21), pp. 6573-6582.

[8] N. Japkowicz, “The class imbalance problem: significance and strategies”, In *Proc. of IC-AI'00*, 2000.

[9] B. Liu, W. S. Lee, P. S. Yu and X. Li, “Partially supervised classification of text documents”, In *Proc. of ICML’02*, 2002, pp. 387-394.

[10] K. Nigam, A. McCallum, S. Thrun and T. Mitchell, “Text classification from labeled and unlabeled documents using EM”, *Mach. Learning*, 2000, vol. 39(2/3), pp. 103-134.

[11] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, C. J. Brown and A. K. Dunker, “Predicting intrinsic disorder from amino acid sequence”, *Proteins*, *Special Issue on CASP5*, in press.

[12] J. C. Platt, “Probabilistic outputs for support vector machines and comparison to regularized likelihood methods”, In *Advances in Large Margin Classifiers*, A. J. Smola, P. Bartlett, B. Scholkopf, D. Schuurmans (eds.): MIT Press, 1999, pp. 61-74.

[13] M. F. Porter, “An algorithm for suffix stripping”, *Program*, 1980, vol. 14(3), pp. 130-137.

[14] F. Provost and P. Domingos, “Well-trained PETs: Improving probability estimation trees”, CeDER Working Paper #IS-0004, Stern School of Business, New York University, 2000.

[15] M. D. Richard and R. P. Lippmann, “Neural network classifiers estimate Bayesian a posteriori probabilities”, *Neural Comput.*, 1991, vol. 3, pp. 461-483.

[16] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval”, *Inf. Process. Manage.*, 1988, vol. 24(5), pp. 513-523.

[17] M. Seeger, “Learning with labeled and unlabeled data”, Technical Report, Institute for Adaptive and Neural Computation, University of Edinburgh, 2001.

[18] B. Shahshahani and D. Landgrebe, “The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon”, *IEEE Trans. Geosci. Remote Sens.*, 1994, vol. 32(5), pp. 1087-1095.

[19] D. M. J. Tax and R. P. W. Duin, “Uniform Object Generation for Optimizing One-class Classifiers”, *J. Mach. Learn. Res., Special Issue on Kernel Methods*, 2002, vol. 2(2), pp. 155-173.

[20] V. N. Vapnik, *Statistical Learning Theory*, Wiley, 1998.

[21] S. Vucetic, D. Pokrajac, H. Xie and Z. Obradovic, “Detection of underrepresented biological sequences using class-conditional distribution models”, In *Proc. of SIAM SDM’03*, 2003, pp. 279-283.

[22] Y. Wu, Q. Tian and T. S. Huang, “Integrating unlabeled images for image retrieval based on relevance feedback”, In *Proc. of ICPR'00*, 2000.

[23] G. Yona, N. Linial and M. Linial, “ProtoMap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space”, *Proteins*, 1999, vol. 37, pp. 360-378.

[24] B. Zadrozny and C. Elkan, “Transforming classifier scores into accurate multiclass probability estimates”, In *Proc. of KDD’02*, 2002, pp. 694-699.