Correcting Sampling Bias in Structural Genomics through Iterative Selection of Underrepresented Targets

Kang Peng Slobodan Vucetic Zoran Obradovic*

Abstract

In this study we proposed an iterative procedure for correcting sampling bias in labeled datasets for supervised learning applications. Given a much larger and unbiased unlabeled dataset, our approach relies on training contrast classifiers to iteratively select unlabeled examples most highly underrepresented in the labeled dataset. Once labeled, these examples could greatly reduce the sampling bias present in the labeled dataset. Unlike active learning methods, the actual labeling is not necessary in order to determine the most appropriate sampling schedule. The proposed procedure was applied on an important bioinformatics problem of prioritizing protein targets for structural genomics projects. We show that the procedure is capable of identifying protein targets that are underrepresented in current protein structure database, the Protein Data Bank (PDB). We argue that these proteins should be given higher priorities for experimental structural characterization to achieve faster sampling bias reduction in current PDB and make it more representative of the protein space.

1 Introduction.

In data mining and machine learning it is commonly assumed that the training, or labeled, dataset is unbiased, i.e., a random sample from the same underlying distribution as the data on which the learned model will be applied. In real-life applications, however, this assumption is often violated due to the *sampling bias* or *sample selection bias* problem [4], and the labeled dataset is no longer a representative of the whole population. Consequently, a predictor model learned on such data may be suboptimal and will not generalize well on out-of-sample examples. Thus, it is crucial to be able to detect and correct such bias, and bias detection and correction should be considered an integral part of the learning process.

The sampling bias can be quantified by the *sample selection probability* $p(s{=}1|\mathbf{x}, \mathrm{y})$ [11], where $\mathbf{x}$ is feature vector, y is target variable (label) and s is a binary random variable. If $s = 1$, an example $(\mathbf{x}, \mathrm{y})$ is selected into the labeled dataset, while if $s = 0$, it is not selected. Thus, if $p(s{=}1|\mathbf{x}, \mathrm{y})$ is constant for all $(\mathbf{x}, \mathrm{y})$, the labeled dataset is unbiased; otherwise, it is biased. In numerous real-life scenarios, a reasonable and realistic assumption is that s is dependent on $\mathbf{x}$ but conditionally independent of y given $\mathbf{x}$, i.e. $p(s{=}1|\mathbf{x}, \mathrm{y}) = p(s{=}1|\mathbf{x})$.

A suitable tool for detecting and characterizing the sampling bias are contrast classifiers [9] that measure the distributional difference between the labeled dataset and an unlabeled dataset, which is a large, unbiased example of the underlying distribution. The contrast classifier is a 2-class classifier trained to discriminate between examples of labeled and unlabeled datasets. Its output is directly related to the sample selection probability $p(s{=}1|\mathbf{x})$ and thus could be a useful measure of sampling bias, as illustrated on a synthetic dataset as well as in real-life bioinformatics applications [9], [8].

This study was motivated by an important bioinformatics problem that is directly related to sampling bias. It is well-known that three dimensional (3-D) structure of a protein is a crucial determinant of its biological function and biochemical properties. However, only a small fraction of known proteins has experimentally determined 3-D structure; currently, there are about 25,000 protein structures deposited in the main protein structure database, Protein Data Bank (PDB) [1], as compared to more than 1.5 million known proteins. Furthermore, current content of PDB is highly biased in the sense that it does not adequately sample the protein sequence space, due to various issues related to experimental structure determination [8].

The ongoing structural genomics projects [2] try to address this problem by experimentally determining structures of a carefully selected set of representative protein targets which, along with those already in PDB, could achieve maximal coverage of the protein sequence space. However, most of the existing target selection strategies [5] rely on sequence comparison methods and, as a result, produce long lists of representative proteins without structurally characterized homologues (i.e. proteins evolved from the same ancestor that typically have similar sequence and structure). It is evident that, using the available technology, the

*Correspondence to Zoran Obradovic, zoran@ist.temple.edu, Center for Information Science and Technology, Temple University, Philadelphia, PA.

progress of these projects is likely to be slow. Thus, in order to rapidly achieve a good coverage of the protein space, novel computational methods for prioritization of protein targets should be developed.

In such application, labeled examples are proteins with experimentally determined 3-D structures. A protein is also considered as labeled if at least one of its homologues has known structure. Unlabeled examples are defined as all known proteins. Thus stated, our specific goal is developing methodologies to select the most informative unlabeled proteins for labeling, i.e. experimental structure characterization, as to rapidly reduce the sampling bias existing in the labeled proteins in PDB.

In this study we propose an iterative procedure for the prioritization based on the contrast classifier framework. Starting from a biased labeled dataset and an unbiased unlabeled dataset, the procedure iteratively builds contrast classifiers to (a) determine if sampling bias in current labeled dataset is significant, and (b) select a certain number of underrepresented unlabeled examples based on the contrast classifier output. The selected examples are then assumed as labeled and added into the current labeled data, thus reducing the sampling bias.

The proposed procedure was applied on a complete genome whose protein sequences are currently used as structural genomics targets. We argue that the proposed approach is highly suitable for prioritizing structural genomics protein targets since it emphasizes importance of the most underrepresented protein sequences. Revealing structural properties of such proteins is likely to produce highly significant biological results.

2 Methodology.

2.1 Problem formulation. Given two datasets sampled from the same unknown distribution, where D_L is *labeled* and *biased*, while D_U is *unlabeled* and *unbiased*, our objective is to identify a set of G most informative examples from D_U which, once labeled, would maximally reduce the labeled data bias. Here G is a user-specified parameter that depends on total labeling cost. We assume that labeling cost is uniform for all examples.

2.2 Contrast classifier for detecting sampling bias. As shown in [9] the contrast classifier is a 2-class classifier trained to learn the distributional difference between labeled and unlabeled datasets. When using classification algorithms that estimate posterior conditional class probability and balanced training data, the contrast classifier output $cc(\mathbf{x})$ approximates $u(\mathbf{x})/(u(\mathbf{x})+l(\mathbf{x}))$, or, equivalently, $l(\mathbf{x})/u(\mathbf{x}) = (1\text{-}cc(\mathbf{x}))/cc(\mathbf{x})$, where $l(\mathbf{x})$ and $u(\mathbf{x})$ are probability density functions (pdfs) for D_L and D_U respectively.

It is straightforward to show the connection between $cc(\mathbf{x})$ and the sample selection probability $p(s{=}1|\mathbf{x})$ [11], where s is a binary random variable indicating whether $\mathbf{x}$ is sampled into the labeled dataset D_L. If $p(s{=}1|\mathbf{x})$ is constant for all $\mathbf{x}$, there is no sampling bias; otherwise, the labeled dataset is biased. Following Bayes theorem, $p(s{=}1|\mathbf{x}) = p(s{=}1)p(\mathbf{x}|s{=}1)/p(\mathbf{x})$, where $p(\mathbf{x}|s{=}1)$ and $p(\mathbf{x})$ can be approximated by $l(\mathbf{x})$ and $u(\mathbf{x})$, respectively. Thus, $p(s{=}1|\mathbf{x}) = p(s{=}1)l(\mathbf{x})/u(\mathbf{x}) = p(s{=}1)(1\text{-}cc(\mathbf{x}))/cc(\mathbf{x})$.

The contrast classifier output could therefore be a useful measure of sampling bias. If $cc(\mathbf{x}) < 0.5$, then $l(\mathbf{x}) > u(\mathbf{x})$ and x is *overrepresented* in D_L. If $cc(\mathbf{x}) > 0.5$, then $l(\mathbf{x}) < u(\mathbf{x})$ and $\mathbf{x}$ is *underrepresented* in D_L. Otherwise, $l(\mathbf{x}) = u(\mathbf{x})$ and $\mathbf{x}$ is equally represented in D_L and D_U. Thus, the $cc(\mathbf{x})$ distribution for D_U could reveal the overall level of sampling bias: the bias is negligible if it is concentrated around 0.5, otherwise the bias is significant. Alternatively, the difference between the two $cc(\mathbf{x})$ distributions for D_L and D_U could also be used to measure the overall bias.

2.3 An iterative procedure for correcting sampling bias. Based on the discussion above, we propose to use contrast classifier output as criterion to select underrepresented examples for labeling to correct sampling bias. A one-step approach for this purpose would be building a single contrast classifier from the initial D_L and D_U and selecting G underrepresented examples from D_U. An open question is what G underrepresented examples would be the most suitable for selection. A one-step approach for selection may be too aggressive and fail to properly correct the bias. Therefore, a challenging problem is how to determine an appropriate selection schedule that would minimize bias of the resulting labeled dataset.

We propose a procedure (Figure 1) that iteratively builds contrast classifiers and incrementally selects underrepresented examples. At each iteration, a contrast classifier is built from current D_L and D_U and then applied to D_U. If the $cc(\mathbf{x})$ distribution for D_U indicates the overall bias is significant, a set of B underrepresented examples will be selected and added into D_L for building the contrast classifier at the next iteration. Details about how to select the B examples will be discussed in the next section. The whole procedure iterates until the required G examples have been selected or the sampling bias becomes negligible.

It should be noted that the actual labeling is not really necessary during the procedure since the label

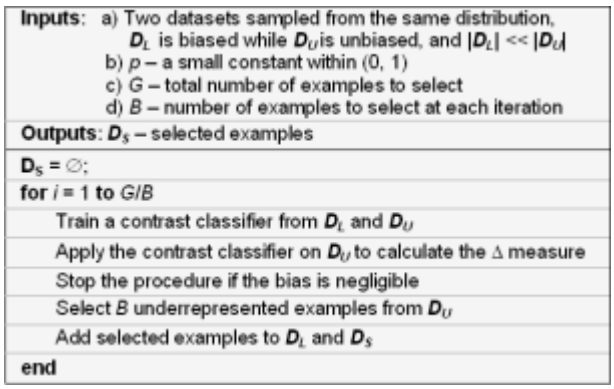

Figure 1: The proposed iterative procedure.

information is not used. It can be done after the procedure stops for all selected examples (D_S), which is the differential set between the final and initial D_L. This is one of the major differences between the proposed procedure and active learning methods [3].

2.4 Selection of underrepresented examples. After a contrast classifier is built and significant bias is detected, potentially underrepresented examples can be selected from D_U. A straightforward method (named *Top-B*) is to simply select B examples with the highest $cc(\mathbf{x})$ from D_U. However, this approach might not be efficient in reducing sampling bias, since these examples may be redundant. They may come from a same underrepresented region that produces similarly high $cc(\mathbf{x})$ values.

An alternative method (named *Random-B*) first determines a threshold θ^p such that only $100p\%$ of labeled examples have $cc(\mathbf{x})$ values higher than it, where p is a small constant in (0, 1). Then, all unlabeled examples that satisfy $cc(\mathbf{x}) > \theta^p$ are considered as underrepresented, i.e. $U^p = \{\mathbf{x} \mid cc(\mathbf{x}) > \theta^p \wedge \mathbf{x} \in D_U\}$. Finally, B examples are randomly drawn from U^p according to the uniform distribution. In this way the selected examples could cover the underrepresented regions more evenly.

2.5 Quantitative measure of overall bias. As discussed in § 2.2, the sampling bias can be assessed qualitatively by visual inspection of the $cc(\mathbf{x})$ distributions for D_L and D_U, i.e. whether the distribution for D_U is concentrated around 0.5, or whether the two distributions are largely overlapped. We also defined a quantitative measure of overall bias $\Delta = sqrt(\sum_{i=1}^{|D_U|}(cc(\mathbf{x}_i) - 0.5)^2/|D_U|)$, where $cc(\mathbf{x}_i)$ is contrast classifier output for the i-th example in D_U. The value of Δ will approach 0 when bias is negligible since all $cc(\mathbf{x}_i)$ should be close to 0.5. It will be large if the bias is significant and many $cc(\mathbf{x}_i)$ are far away from 0.5. In the extreme case when all $cc(\mathbf{x}) = 1$, $\Delta = 0.5$.

3 Bioinformatics Application in Structural Genomics.

In this section we applied the proposed procedure to prioritizing structural genomics targets from a model organism. We show that the proposed iterative procedure is more effective than simple random sampling and a one-step approach discussed in § 2.3.

3.1 Datasets. Since the number of known proteins is large, we limited our study to the 28,334 protein sequences (40 amino acids or longer) from a model organism called *Arabidopsis thaliana* (ATH) extensively studied in plant biology. It is also a major source of structural genomics targets for the Center for Eukaryotic Structural Genomics (CESG, http://www.uwstructuralgenomics.org/).By now CESG has selected about 4,000 target proteins from this genome and finished structure determination for 19 of them.

The amino acid sequences were obtained from website of the Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org/). As in a previous study [8], a non-redundant representative subset of 14,988 sequences was selected, with no two sequences having pairwise identity higher than 40%. Out of these sequences 838 were identified to have known structures and thus formed D_L, while all of the 14,988 proteins were assigned to D_U. We assume that at most 500 proteins, i.e. $G = 500$, can be selected from D_U for labeling due to available resources for experimental structure determination.

3.2 Knowledge representation and contrast classifier training. A similar knowledge representation was adopted as in the previous study [8], i.e. constructing one example per sequence position instead of one example per sequence. Each example consisted of 30 attributes and a class label of 0 or 1 indicating if it was from D_L or D_U. In addition to the 25 attributes used in the previous study [8], 3 transmembrane helix predictions by PHDHtm predictor [10], 1 disorder prediction by VL3 predictor [7] and 1 coiled-coils prediction by COILS predictor [6] were also included.

The contrast classifiers were built as an ensemble of 20 neural networks each having 10 hidden neurons and 1 output neuron with sigmoid activation function. A two-stage sampling procedure [8] was employed to construct balanced training sets (12,000 examples) for training component networks. The contrast classifiers

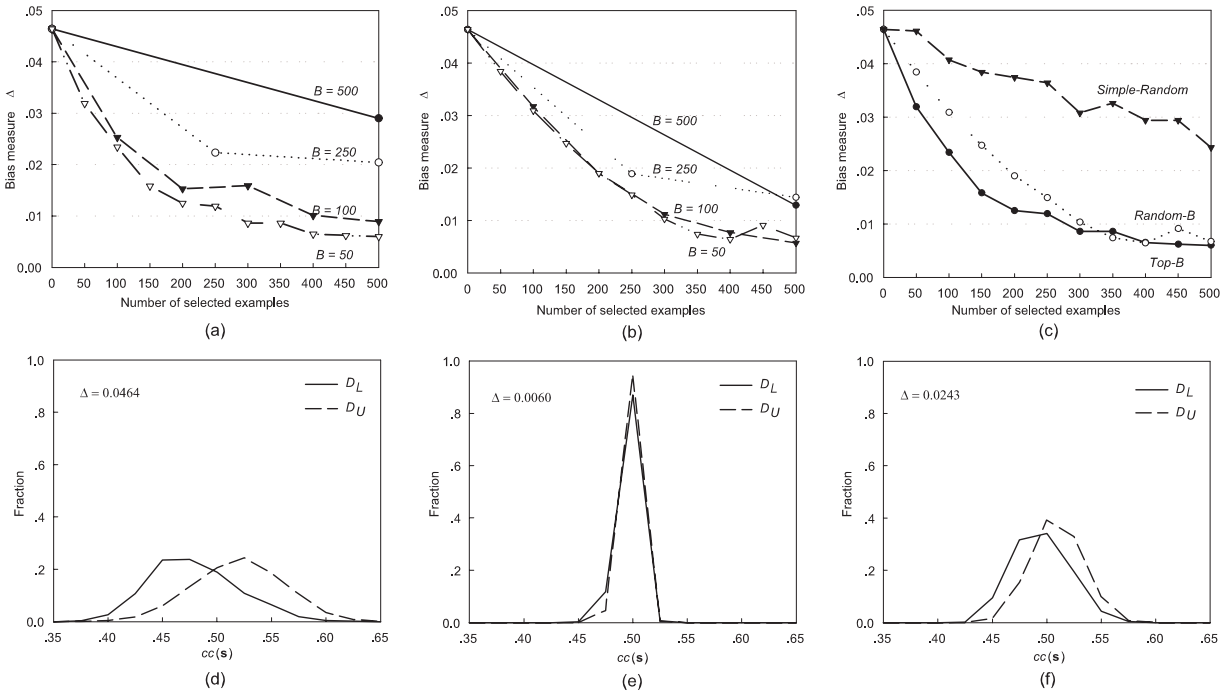

Figure 2: Application of the proposed procedure on structural genomics data. Upper row - plots of Δ vs. *Number of examples selected* for (a) *Top-B*, (b) *Random-B* and (c) Comparison to *Simple-Random*. Lower row - $cc(\mathbf{s})$ distributions for D_L and D_U (d) at the beginning, after 500 proteins selected using (e) *Top-B*($B = 50$), and (f) *Simple-Random*. The distributions for *Random-B* are similar to those of *Top-B* and thus not shown.

were applied to each sequence $\mathbf{s}$ in D_U to obtain $cc(\mathbf{x})$ for each sequence position. These *per-position* predictions were then averaged over the sequence as the *per-sequence* prediction $cc(\mathbf{s})$, which was used to select underrepresented sequences.

3.3 Application of the proposed procedure. We examined different values of $B = \{50, 100, 250, 500\}$ for the proposed procedure. Note that $B = 500$ corresponds to the one-step approach of selecting all 500 proteins at one time (§ 2.3). The two methods *Top-B* and *Random-B* ($p = 0.05$) for selecting underrepresented examples were compared. The performance was assessed using the Δ measure of overall bias (§ 2.5) calculated in each iteration. A successful bias correction procedure should result in a rapid decrease in Δ measure as more proteins are selected. In Figure 2a and 2b we show the plots of Δ measure *vs.* number of selected proteins for the two methods.

As evident from the plots, the proposed iterative procedure ($B = 50$) performed better than the one-step approach ($B = 500$) in the sense that it achieved much lower Δ value, or lower level of overall bias, after selected 500 proteins. However, the difference is smaller for *Random-B* than for *Top-B*. It is worth noting that smaller B typically leads to better bias reduction but with larger computational efforts. However, the small difference between the plots for $B = 50$ and $B = 100$ indicates that the gain of using even smaller B may be marginal.

In addition to the proposed procedure, we also examined the simple approach (*Simple-Random*) of randomly selecting M unlabeled proteins in a single step. For each $M = \{50, 100, 150, \ldots, 500\}$, a contrast classifier was built to calculate the Δ measure of overall level of bias after the M proteins were added into D_L. The plot of Δ measure vs. number (M) of selected proteins is shown in Figure 2c, along with those for the proposed procedure ($B = 50$). As expected, the *Simple-Random* method was less effective in bias reduction, with the final $\Delta = 0.0243$, as compared to 0.0060 for *Top-B* and 0.0067 for *Random-B*. This is further confirmed by the $cc(\mathbf{s})$ distributions in Figure 2. The two resulting distributions are almost completely

overlapped for *Top-B* with $B = 50$ (Figure 2e), but still clearly separated for *Simple-Random* (Figure 2f), after 500 proteins were selected.

Out of the 500 proteins selected using *Top-B* (*Random-B*) method, only 68 (79) are currently selected as structural genomics targets by the Center for Eukaryotic Structural Genomics and none of them have been solved. We argue that the remaining 432 (421) proteins should also be selected as structural genomics targets and should be given higher priorities. Along with proteins with known structures, these proteins should be very helpful in achieving maximal coverage of the protein space for *Arabidopsis thaliana* genome.

4 Conclusions.

In this study we proposed an iterative procedure for correcting sampling bias in labeled data. This approach is applicable if an unbiased unlabeled dataset is available. It iteratively builds contrast classifiers to detect sampling bias and selects underrepresented examples which, once labeled, can be very effective in correcting the sampling bias in labeled data. As illustrated on an important bioinformatics problem of prioritizing protein targets for structural genomics projects, the proposed procedure is more effective than randomly choosing unlabeled proteins for labeling in reducing sampling bias, and can rapidly achieve a good coverage of the protein space.

The ultimate solution for correcting sampling bias is to add new labeled examples. A simple approach is to randomly select examples for labeling. As more examples are labeled, the bias in the resulting labeled dataset would be gradually reduced. This is especially good if the total number of new labeled examples could be relatively large. However, in real-life problems like protein structure determination it is often the case that the costs for labeling even a single example could be very high. Consequently, only a small number of examples can be allowed and thus the selected ones should be the most informative. In such scenarios, our procedure could be very appropriate since it emphasizes the underrepresented examples based on contrast classifier output directly related to the sample selection probability $p(s{=}1|\mathbf{x})$.

More work is needed to fully characterize the proposed procedure. An improved performance measure is needed since the Δ measure may depend on the learning algorithms used in learning contrast classifiers. As shown in § 3.3, smaller B might be desirable for better bias reduction but would require more computational efforts. Thus, additional analytical and experimental work needs to be done to determine the optimal tradeoff between the computational effort and the level of bias reduction, given a fixed total number G of allowed examples for labeling. Finally, if G is very small, it is likely that more elaborate procedure would be needed than the proposed *Top-B* and *Random-B* procedures. These issues are the subjects of our ongoing research.

Acknowledgements. We thank A. K. Dunker, director of Center for Computational Biology and Bioinformatics at Indiana University School of Medicine, for pointing us to structural targets prioritization in *Arabidopsis thaliana* genome as a potentially high impact application domain for our sampling bias correction method. We are also grateful to J. L. Markley, director of the Center for Eukaryotic Structural Genomics at University of Wisconsin-Madison, for providing us access to CESG structural genomics target database for a more detailed biological relevance validation of our method (results not included in this report).

References

[1] H. M. Berman, T. N. Bhat, P. Bourne, Z. Feng, G. Gilliand and H. Weissig, *The Protein Data Bank and the challenge of structural genomics*, Nat. Struct. Biol., 7 (2000), pp. 957–959.

[2] S. E. Brenner, *A tour of structural genomics*, Nat. Rev. Genet., 2 (2001), pp. 801–809.

[3] D. Cohn, L. Atlas and R. Ladner, *Improved generalization with active learning*, Mach. Learning, 15 (1994), pp. 201–221.

[4] J. Heckman, *Sample selection bias as a specification error*, Econometrica, 47 (1979), pp. 153–161.

[5] M. Linial and G. Yona, *Methodologies for target selection in structural genomics*, Prog. Biophys. Mol. Biol., 73 (2000), pp. 297–320.

[6] A. Lupas, M. Van Dyke and J. Stock, *Predicting coiled coils from protein sequences*, Science, 252 (1991), pp. 1162–1164.

[7] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, C. J. Brown and A. K. Dunker, *Predicting intrinsic disorder from amino acid sequence*, Proteins, 53 Suppl 6 (2003), pp. 566–572.

[8] K. Peng, Z. Obradovic and S. Vucetic, *Exploring bias in the Protein Data Bank using contrast classifiers*, Pacific Symposium on Biocomputing, Hawaii, (2004), pp. 435–446.

[9] K. Peng, S. Vucetic, B. Han, H. Xie and Z. Obradovic, *Exploiting unlabeled data for improving accuracy of predictive data mining*, 3rd IEEE Int'l Conf. on Data Mining, Melbourne, FL, (2003), pp. 267–274.

[10] B. Rost, R. Casadio, P. Fariselli and C. Sander, *Prediction of helical transmembrane segments at 95% accuracy*, Protein Sci., 4 (1995), pp. 521–533.

[11] B. Zadrozny, *Learning and evaluating classifiers under sample selection bias*, in C. E. Brodley, ed., 21st Int'l Conf. Machine Learning, ACM Press, Banff, Alberta, Canada, (2004).