Generalization Methods in Bioinformatics

Steven Eschrich
Department of Computer Science and Engineering
University of South Florida
4202 East Fowler Avenue,
Tampa, FL 33620
eschrich@csee.usf.edu

Nitesh V. Chawla
Department of Computer Science and Engineering
University of South Florida
4202 East Fowler Avenue,
Tampa, FL 33620
chawla@csee.usf.edu

Lawrence O. Hall
Department of Computer Science and Engineering
University of South Florida
4202 East Fowler Avenue,
Tampa, FL 33620
hall@csee.usf.edu

ABSTRACT

Protein secondary structure prediction and high-throughput drug screen data mining are two important applications in bioinformatics. The data is represented in sparse feature spaces and can be unrepresentative of future data. Supervised learners in this context will display their inherent bias toward certain solutions, generally solutions that fit the training set well. In this paper, we first describe an ensemble approach using subsampling that scales well with dataset size. A sufficient number of ensemble members using subsamples of the data can yield a more accurate classifier than a single classifier using the entire dataset. Experiments on several datasets demonstrate the effectiveness of the approach. We report results from the KDD Cup 2001 drug discovery dataset in which our approach yields a higher weighted accuracy than the winning entry. We then extend our ensemble approach to create an over-generalized classifier for prediction by reducing the individual subsample size. The ensemble strategy using small subsamples has the effect of averaging over a wider range of hypotheses. We show that both protein secondary structure prediction and drug discovery prediction can be improved by the use of over-generalization, specifically through the use of ensembles of small subsamples.

1. INTRODUCTION

Bioinformatics is a rapidly expanding field involving a significant contribution from the data mining community. Many different problems are grouped into the general category of bioinformatics, including protein secondary structure prediction, drug discovery, DNA microarray analysis, gene prediction and genome analysis [17]. In this paper, we focus on two particular problems in which data mining can contribute: drug discovery and protein secondary structure prediction. Both of these fields demonstrate distinct characteristics as data mining problems. We will discuss the two areas and then attempt to describe what we believe is the common thread — the need for "over-generalization" on the part of any learner. Once we have described the overall concept we introduce a subsampling technique that we believe demonstrates this generalization. We also show that predictive accuracy can be improved through the use of over-generalization.

1.1 Protein Secondary Structure Prediction

Proteins are a key component of life and central to our understanding of cellular processes. A protein consists of a linear chain of amino acids; each amino acid can be one of twenty different types. The key nature of proteins is not their one-dimensional structure, but rather how they fold back on themselves. Proteins can take on several different configurations and these are chemically important. Many different forces act upon the protein chain in order to create such folds, including ionic, hydrogen, van der Waals and hydrophobic interactions [13]. Much of the process by which a sequence of amino acids forms a more complex structure is not completely understood. X-ray crystallography has been the principal means of elucidating the structure of many proteins, however it has been applied to only a small fraction of the total number that can exist [13]. The process is difficult and error-prone. However, if we wish to understand a cellular process or perhaps design a drug to inhibit a protein-ligand interaction we must first understand the secondary structure of a protein. Presently structural prediction is most successful when sufficiently similar homologs are known [17]. Homologs, or similar protein chains, can be used to estimate the structure of an un-visualized protein. When there are no close homologs to a protein, the structural prediction can be very poor.

New protein sequences are constantly being discovered and often comprise very different and distinct chains of amino acids. Modern biochemistry can not yet fully describe the ways in which a strand of amino acids may fold. Data mining techniques can be used in an attempt to predict this structure. The CASP contest [21] is one method of encouraging the research community to focus on secondary structure prediction. In the latest such contest (CASP-4), the winner was PSIPRED [18]. The author used the PSI-BLAST [1] profiles for a set of protein chains as training data for a neural network. The network used a window of 15 amino acids within the protein chains and an ensemble of four neural networks was built. Each feature was the log likelihood of substitution of an amino acid. Earlier work by some of the authors have attempted to understand the process through which PSIPRED succeeded. We provide one such explanation here, both as a followup and a new direction from earlier work.

1.2 Drug Discovery

The second problem we consider within this paper is drug discovery. This is an challenging field that encompasses both

protein secondary structure prediction and many other difficult steps. For this paper, we consider only one small aspect of the entire process — prediction of drug-like activity of a compound based on its structure. We use a simplified description of a molecule and attempt to predict if the molecule will exhibit drug-like behavior in a high-throughput screening. Drug discovery encompasses understanding cellular processes, predicting protein structures, and estimating interactions between a molecule and the normal biological molecular targets [12]. Chemists and biologists would ideally like to fully understand the pathways involved in a disease and from this knowledge develop a molecule (or several molecules) that can interact with the disease agents to neutralize them [12]. However, many complex interactions are occurring at the cellular level that makes the full rational drug design process extremely difficult. Modern pharmaceutical companies pursue methods that attempt to circumvent some of these problems through the use of high-throughput screening. Since we do not have full knowledge of the physical and electrostatic forces governing complex interactions of proteins, enzymes and ligands, we can instead try to test many available compounds for a desired response. This is the so-called lead finding stage of drug discovery. Modern testing technology has developed to the point that this process can be automated. Many compounds can be quickly tested and leads can be identified from these results. These leads can be analyzed more closely and structurally similar compounds can be investigated for potentially higher efficacy.

Data mining within drug discovery involves predicting if a compound is likely to demonstrate drug-like activity in the presence of a given disease (or simply a given chemical target). High-throughput screens result in a large amount of data for analysis. Ideally, in silico screening of compounds could replace the need for HTS in identifying leads. Current data mining techniques can not only be used to predict an untested compound's activity but also can identify inconsistent and potentially incorrect HTS results. We use data mining tools to predict the activity of a molecule based on solely on it's structural characteristics. Much work in drug discovery involves the hypothesis that compounds with similar structure are likely to exhibit similar pharmacological activity [24; 7]. We use the simple bit-string fingerprint representation of a compound's structure [11; 28]. The atom-pair fingerprints were generated by Tripos, Inc.

2. BACKGROUND

2.1 Ensembles

An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way to classify new examples [14]. Many popular meta-learning techniques in computer science can be conveniently described within this framework. An ensemble consists of a set of possibly different classifier types. The output of each classifier is combined in one of many different ways in order to reach a final classification. This definition is broad, however it encompasses many different popular techniques within the same framework, including bagging [4] and boosting [16].

The approach introduced within this paper is similar in spirit to bagging [4]. Bagging, or bootstrap aggregation, generates many bootstrap samples of a dataset, using sampling with replacement. Each sample is the size of the training set, although it will contain multiple copies of some examples and not include others. For sufficiently large datasets, it can be easily shown that these samples will contain approximately 63.2% of the unique examples from the original dataset [2]. The sample is then used to construct a model to be used by the classifier. The overall bagged classifier (ensemble) is the aggregation of each individual classifier. The outputs are combined via majority vote, a common and reasonable approach when using ensembles. Breiman also introduced the notion of Rvotes [6], which is an ensemble of classifiers similar to bagging. However, Rvoting uses subsamples that are much smaller than the full size of the training set. Results in [6] indicate the approach is not always as effective as bagging.

Empirical evaluations of ensembles show that they often outperform individual classifiers. Several requirements are generally considered necessary:

1. Ensemble members must be diverse.
2. Ensemble members must be reasonably accurate.

Diversity in ensembles is a crucial requirement. Dietterich [14] argues this can be seen as the averaging of many different consistent hypotheses within the same region of hypothesis space. Statistics and common sense tell us that with no other knowledge, taking the average of the possible estimates for a value yields a reasonable approximation to the value. Specifically, this averaging tends to reduce the variance introduced by individual models [20]. This argument also indicates the rationale behind the requirement that ensemble members must be reasonably accurate. Otherwise we include many outliers or inconsistent hypotheses in the averaging operation, thus increasing the variance of the averaged hypothesis. The use of diversity in classification is an important issue within our approach; we propose generating many diverse classifiers.

2.2 Subsampling and Scalability

Subsampling is a popular technique for data reduction in data mining and machine learning. Many different approaches are possible and statistical bounds on worst-case error can be used to determine a worst-case subsample size. On the other hand a purely empirical method would require plotting an error curve using increasingly larger subsample sizes until a plateau in error is seen [26]. The progressive sampling technique addresses the determination of sample size. However progressive sampling requires results from prior iterations thereby creating a serial learning approach. Single subsamples, in either the empirical or theoretical approach, are often widely varying in classification accuracy depending on the particular random subsample. For instance, the accuracy from successive subsamples of size 50% can widely vary from subsample to subsample. Ensemble methods are successful at averaging classification results from the ensemble members. This is the strategy used by bagging [4]. Many diverse classifiers can be generated and the overall ensemble is in some sense an average of the representative region of hypothesis space. Moreover, ensemble methods like bagging are parallel in nature. Each bag can be generated and a model can be constructed in parallel. Bagging is not a scalable approach to data mining since each individual bag is the size of the full training set. Therefore, we would like

to combine the use of subsampling with the averaging (and generalization) of ensembles.

Some of the authors previously addressed the inadequacy of using small bags in a protein secondary structure prediction problem [9]. They concluded that choosing random, disjoint partitions of a dataset creates ensembles that are as accurate as using the same size subsamples (with or without replacement). Generating bags of very large datasets are believed to potentially create overly similar models, thus hurting the overall ensemble. Below we intend to address this issue and describe some empirically determined heuristics for determining appropriate parameters for subsampling.

3. APPROACH

Protein secondary structure prediction and drug discovery data mining are both difficult problems in bioinformatics. They have several common features with respect to data mining. First, the data comes from very noisy environments. Protein structures are not always completely correct and thus a training set of structures may consist of possibly inconsistent results. High throughput screening is a very noisy environment in which the actual measurement of activity can be disturbed by contaminants and environmental conditions. Noisy data is nothing new to the data mining community, however the total problem space is sufficiently large and the number of labeled examples is sufficiently small that noise can be a significant problem.

Another common difficulty in both structure prediction and drug discovery involves the violation of the iid assumption. This assumption is that each example is independently and identically distributed. Machine learning uses this reasonable assumption as the basis for creating generalized models from a set of examples. Otherwise, a classifier using these models is constructed to predict properties for which nothing is known or can be inferred. However, the independence assumption is generally weak for both protein structure prediction and drug discovery. Protein secondary structure prediction is necessarily limited to those proteins that have been investigated and submitted to structural analysis. Predictions are desired for proteins that may be significantly different in structure, for example new and novel sequences. Drug discovery also involves this issue — candidate compounds may be screened because they are likely to be successful or because they are the most plentiful in supply. For instance, the Diversity set of compounds from the National Cancer Institute is composed of diverse chemical representatives, however the set of choices was necessarily restricted to those for which there was a sufficient supply [25].

Classifier generalization is the ability of a classifier to make reasonable predictions beyond the data it has already seen. If a classifier simply memorizes the existing patterns, then nothing can be said about new examples. Thus all classifiers generalize to some extent. We propose the use of "over-generalization", or generalizing the classifier response as broadly as possible to avoid the hazards inherent to these two domains. This principle is used in bagging [4]; we introduce a similar technique for over-generalization in a scalable framework. In particular, Dietterich [15] observed that bagging with C4.5 generates very accurate classifiers, yet the classifiers are not very diverse (as compared to boosting or randomized C4.5). By using less data (through small subsamples) we expect to create less accurate individual classifiers [19]. The addition of C4.5, an unstable learner, yields an ensemble of weak, diverse classifiers. Stability in a classifier can be achieved through averaging over many individual instances [5]. Since we increase the diversity of the group of classifiers, we expect more classifiers to be needed to achieve stability in the ensemble.

Therefore, we expect that an ensemble of subsamples can produce a stable and accurate composite classifier. Breiman's bagging and Rvote schemes are the two extremes in ensembles of subsamples. This technique is also similar to the work presented in [19] in which very weak models are generated randomly. Many such weak models result in an overall accurate classifier. We empirically investigate a middle-ground in this work, in which the classifiers can be made successively weaker by reducing the subsample size. We determine some heuristics to choose subsample size and the number of ensemble members. We also investigate the ability of over-generalization in each classifier of an ensemble to more accurately predict the non-homologous structures seen within the protein secondary structure prediction dataset.

Several key decisions must be made with regard to the ensemble of subsamples algorithm. Random subsampling of the dataset can be done with or without replacement. Subsampling with replacement is often used due to its simpler implementation and simpler statistical math [23]. When using ensembles of subsamples for over-generalization, we wish to find a small subset of data that can both be representative and minimal. Therefore sampling with replacement would reduce the number of unique representatives within the reduced training set, which could handicap the learner's ability. We examined the results from subsampling with replacement on four modest-sized datasets and found the classification accuracies tended to be lower. We believe that the need for coverage in the training dataset is too important to allow a replicated example to take the place of a potentially important representative example. The bagging technique does not face this problem, since many large bootstrap samples are taken. Minimizing the size of the subsample dictates the use of a no-replacement subsampling strategy.

Another key factor in the ensemble of subsamples approach is the combination strategy for the ensemble. Ensembles are combined using many different schemes, including majority voting, weighted voting based on accuracy estimates and meta-learning over classifier output. Majority voting is often used and is generally a good choice [22]. Therefore we chose a simple majority vote method, both for the simplicity and scalability of the approach.

The implementation of an ensemble of subsamples is straightforward and requires only two parameters: the size of each subsample and the total number of subsamples taken. We restrict the subsample size to the range of $[5\%, 50\%]$ of the size of the original training set for experimentation. We find that the larger sample sizes produce better classifiers than smaller ones. The total number of subsampled sets taken is described below. One key contribution of this paper is the correlation between the subsample size and the number of subsamples.

3.1 Ensemble Task Size

An important design parameter in any ensemble of classifiers is the number of ensemble members. In the context of subsampling, we must decide on the number of independent subsampled sets to create. We can simply choose a

range of numbers (e.g., 1-20 subsamples) and evaluate the different ensembles. However, we find that normalizing the number of subsamples taken relative to the subsampling size creates an easier and more accurate comparison between ensembles using different subsampling percentages. We do this by defining two measures of task size, which are related to the number of examples used in learning.

We first define the size of an individual learning task, LTS. The LTS is the number of examples required for learning, as a fraction of the total dataset size. The size of the total dataset is used as the normalizing factor. A single learner using the entire dataset for learning would have an LTS of 1.

$$LTS = \frac{num\_examples}{total\ dataset\ size} \quad (1)$$

The overall complexity of the ensemble can be measured as the Ensemble Task Size (ETS). With the use of subsampling, each learner is given a reduced number of examples from which to learn. Therefore individual learning tasks are reduced in complexity. However, we would like to capture the notion that the ensemble as a whole may have an increased overall learning task size. The learning can be done in parallel, therefore the ETS does not necessarily indicate longer training time. The ETS is simply derived from the LTS as seen in Equation 2. Two illustrative examples are included below.

$$ETS = \sum_{i=1}^{E} LTS_i \quad (2)$$

Consider first the simple data partitioning strategy of random disjoint partitioning. A dataset is simply split into four equal-sized partitions of $1/4^{th}$ size of the full dataset. Each learner in this simple ensemble would individually have a learning task size (LTS) of 0.25. The computed ETS for this example is 1.0, reflecting the fact that the problem as a whole has not been enlarged or decreased. Next, consider bagging as an ensemble strategy. Each ensemble member is given a bag, or subsample (with replacement) of size equivalent to 100% of the dataset. Therefore each ensemble member has an LTS of 1.0. Thus we do not reduce the size of any individual learner's task. Additionally, the total Ensemble Task Size (ETS) is $ETS = 4 \times 1.0$ or 4.0. This larger value for the ETS represents the fact that we are in some sense increasing the total number of examples learned (across ensemble members) and clearly indicates the scalability problems inherent to bagging.

The ETS measures the complexity of the ensemble in terms of the number of examples in the original dataset. Therefore we can directly compare ensembles of subsamples with different subsample sizes. In addition, we have a direct method of comparing the various subsamples with the well-known bagging approach.

4. RESULTS

We begin by demonstrating the accuracy of small subsamples on several commonly-used machine learning datasets from the UCI Machine Learning Repository [3]. The datasets chosen were page block, pendigits, satimage and letter. The letter dataset was the largest with 20,000 examples. A ten-fold cross-validation was performed for each combination of subsample size and ETS. The accuracy reported is the mean accuracy across the 10 folds. We use C4.5R8 [27] as the learner. Table 1 shows the performance on these reference datasets using C4.5 and small subsamples. Figures 1–4 show accuracy results for these datasets. The choice of using 25% subsamples and an ETS of 5 was observed to be the most accurate. A one-tailed paired-difference t-test was performed between the folds using C4.5 and the small subsamples at the 95% confidence level. In each of the reference datasets, the improvement in accuracy was statistically significant as noted in the table.

Dataset	C4.5	Subsampling 25%, ETS=5	t value
page	96.95	97.28†	2.90
satimage	86.51	89.08†	9.11
pendigits	96.33	97.52†	5.69
letter	88.08	90.44†	9.76

Table 1: UCI Repository Results. †denotes statistical significance.

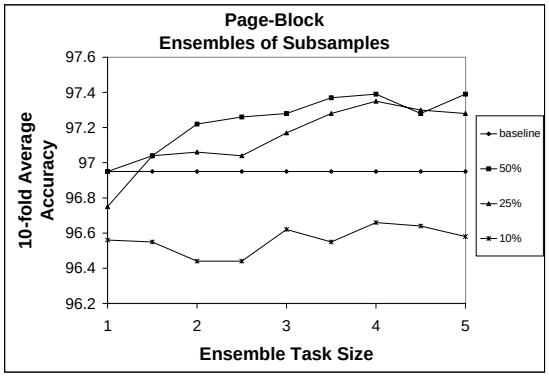

Figure 1: 10-fold results on the page-block dataset. Results are average accuracy across the folds.

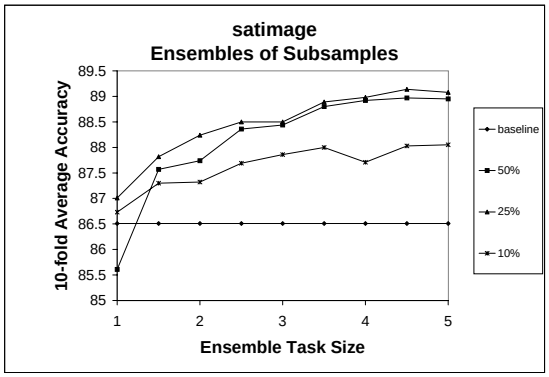

Figure 2: 10-fold results on the satimage dataset. Results are average accuracy across the folds.

We next demonstrate our ensemble of small subsamples approach on both a protein secondary structure prediction

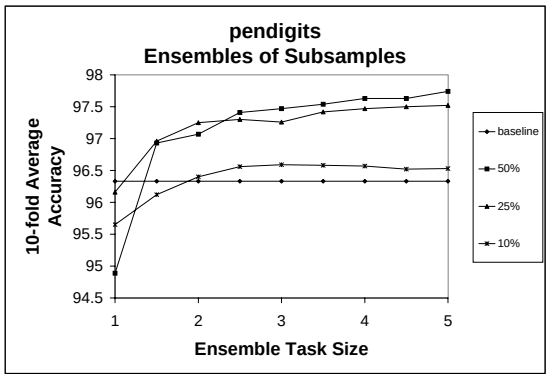

Figure 3: 10-fold results on the pendigits dataset. Results are average accuracy across the folds.

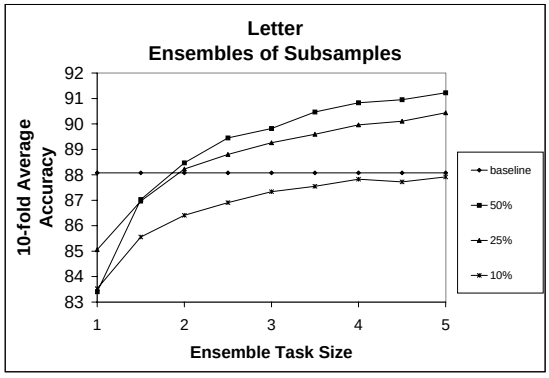

Figure 4: 10-fold results on the letter dataset. Results are average accuracy across the folds.

problem and two drug discovery problems. We again use the C4.5R8 [27] decision tree learner for experiments. Several of the problems use existing training and testing sets; we use ten-fold cross-validation where no such training and testing sets exist. Additionally, we examine the protein secondary structure prediction dataset using both a ten-fold cross-validation on the training set and later using the provided test set.

The MAO (Monoamine Oxidase) inhibitor dataset used in this paper was originally used in [7] for drug activity prediction. It consists of a diverse set of compounds manually selected from the Abbott Laboratories stock of potential drugs. These compounds are represented in a fingerprint format, known as an atom-pair fingerprint and were generated by Tripos, Inc. All pairs of atom types are considered, and for each pair of atom types (for example, nitrogen and carbon) the smallest number of bonds between each N-C pair are calculated. A 10 bit feature represents these distances as either existing or not within the molecule. This process is done for all atom type pairs, generating a bit string of 1200 features (120 pairs of atom types, each pair having 10 bits). The dataset consists of the atom-pair fingerprint and the classification. Each compound was determined to be active or inactive in a particular assay – based on the measured inhibition of MAO.

The KDD Cup 2001 contest consisted, in part, of predicting the activity of compounds binding to thrombin. The representation of the compounds is in an unknown bit-string format, consisting of 139,351 features (bits). Compounds are classified as either active or inactive much like the MAO dataset. There were only 1,909 training examples and only 42 active examples. The test set consisted of 634 compounds. The winner from the KDD Cup 2001 contest detailed his approach in [10]. Surprisingly, he was able to use only four particular features (bits) in his final classifier. We will use the same four-feature dataset from [10]. Accuracies for this contest problem are reported as weighted accuracies, which is the average accuracy computed from the individual class accuracies. The winner was able to achieve a 68.44% weighted accuracy using a Bayesian network [10].

The protein secondary structure prediction dataset is described in [18] as the training and testing sets ("train and test set one") used to develop and validate the neural network that won the CASP-3 secondary structure prediction contest. It consists of a set of 1,219 protein chains. The proteins are submitted to PSI-BLAST [1] and a scoring matrix representing the log-likelihood of each of the amino acids being substituted is returned. This matrix for the protein is split into windows of 15 amino acids and each window is used as an example, together with the protein structure (helix, strand or coil). In addition, an N/C terminus bit is added for amino acid. The training set consists of 209,529 examples of dimension 315, representing 1,156 protein chains. The test set (used in the second set of experiments) consisted of 17,731 examples representing 63 protein chains.

We analyze the effects of ensembles of subsamples on classification accuracy within the two bioinformatics domains described above. We use the MAO dataset and the Jones training set for our first experiments. Figures 5 and 6 show the average 10-fold classification accuracy as both the subsample size and the number of subsamples vary. Recall that the ensemble task size is the number of subsamples (normalized by the size of the subsample). The general trend in both graphs is that the accuracy tends to stabilize at approximately an ETS of 2 to 4. At this point, we see that increasing the subsample size leads to higher accuracies.

For the MAO dataset, only the 50% subsample size performs consistently better than the baseline C4.5. Table 2 shows the ten-fold results for both the baseline C4.5 and the best performing subsample for the MAO dataset. This result is statistically significant in a paired-difference t-test between fold results using a one-tailed test at the 90% confidence level. However, even in the cases in which the improvements are not statistically significant, the classification accuracy on drug activity prediction can still be considered biologically significant. In the Jones training dataset, all but one of the subsamples perform better than the baseline C4.5, with statistical significance. An important observation from these graphs is that in both cases the accuracy improved with larger subsample size. Also, an ensemble in which each learner used at most 50% of the full training data is more accurate than a single model built using the entire dataset.

Next we examined the effect of subsampling on the thrombin dataset. Figure 7 shows the test set weighted accuracy as the subsample size and ETS vary. We see that the 10% subsamples approach is much lower in accuracy than the baseline C4.5 method. Both the 25% and 50% subsample approaches yield higher accuracies than the baseline. In fact, the weighted accuracy using 50% subsamples with an ETS

Baseline C4.5	50% subsamples ETS=4.5
87.88	87.88
86.67	86.67
82.42	85.45
84.76	85.37
82.93	84.76
84.76	84.15
86.59	86.59
86.59	85.98
87.20	87.20
84.76	87.20

Table 2: MAO AFP 10 fold cross-validation comparison.

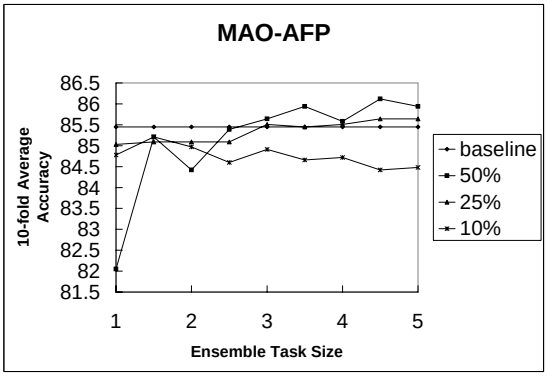

Figure 5: 10-fold results on the MAO afp dataset. Results are average accuracy across the folds.

of both 2 and 3 was 68.55%. The KDD Cup 2001 winning weighted accuracy was 68.44%. Therefore the subsampling approach using 4 features was able to outperform the winning KDD Cup entry.

Finally, we explore the use of over-generalization in two of the datasets. In the case of the MAO dataset, we use a modified version of the fingerprint. This fingerprint still uses the atom-pair concept, however rather than storing the existence of an atom pair at a particular bond length we also include the count of the number of such bonds. Therefore instead of 1200 bits in the representation, we have 1200 integers. This different fingerprint significantly expands the size of the feature space and therefore spreads the existing compounds out in this feature space. We expect it to be more difficult to predict activity as a result. Over-generalization should be important in this context to capture the general relationships among the compounds. We can directly test the over-generalization ability of our ensemble on the Jones dataset by using the training set and the test set, which consists of non-homologous chains. Predicting the structure of non-homologous chains is very difficult for a classifier. We believe that by creating overly-generalized classifiers we can capture some of the large-scale patterns in the data.

Figures 8 and 9 show the over-generalization tests on the protein secondary structure prediction and drug discovery datasets. In both cases, we see the opposite effect as compared to the results from the previous experiment — the smaller subsample size actually increases performance. The non-homologous protein chains require a more general model of protein secondary structure in order to estimate the struc-

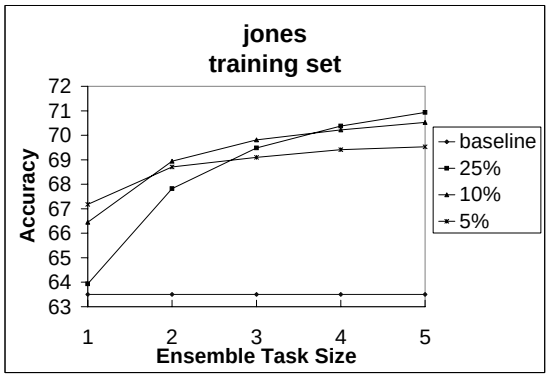

Figure 6: 10-fold results on Jones training set. Results are average accuracy across the folds.

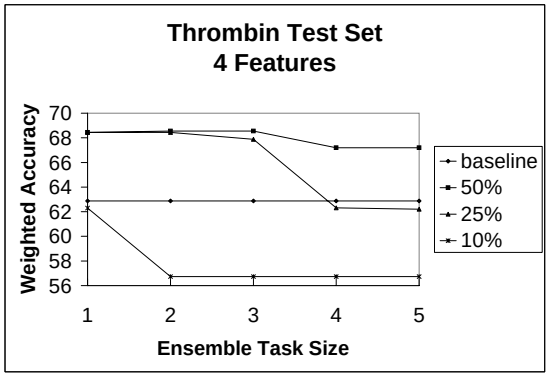

Figure 7: Thrombin (KDD Cup 2001) test set evaluation.

tural differences on such widely different and unseen examples. Likewise for the drug discovery dataset we see that by the use of the atom-fingerprint count representation we are creating a much larger feature space, requiring a much more generalized classifier. Table 3 shows the results for both baseline C4.5 and for 25% subsamples with an ETS of 5. The improvement on the MAO dataset using subsampling, in this case, is statistically significant using a one-tailed paired-difference t-test at the 95% confidence level. Thus, in both cases we force the classifiers to over-generalize and therefore provide better generalization on unseen instances. Generalization occurs in a scalable fashion by taking many small subsamples of the dataset.

Baseline C4.5	25% subsample ETS=5
84.85	87.27
86.67	86.67
83.64	84.85
83.54	84.76
82.93	82.93
79.27	85.37
85.98	87.20
85.98	85.98
85.37	87.20
86.59	87.20

Table 3: MAO AFPC 10 fold cross-validation comparison.

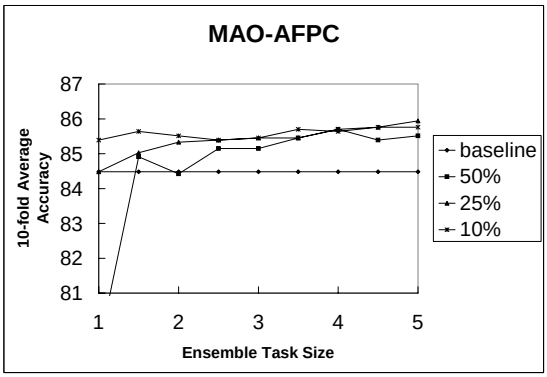

Figure 8: 10-fold cross-validation results on MAO AFPC dataset. Results are average accuracy across the folds.

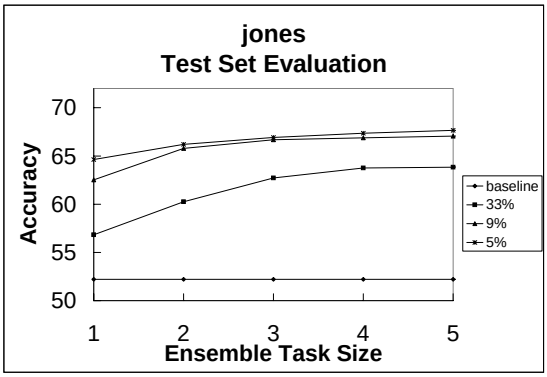

Figure 9: Jones test set evaluation.

5. CONCLUSION

The ensemble of small subsamples approach is a scalable, distributed learning method. The major difficulty involved in subsampling, namely classifier instability, is mitigated using the ensemble approach. The approach is scalable, since no single learner is required to use the entire training set. We empirically show through a series of experiments that the use of 25% subsamples and an ensemble task size of 4 can exceed the performance of C4.5 using the entire training set on three datasets from the bioinformatics domain. An ensemble of subsamples can even exceed the winning classification accuracy on the KDD Cup 2001 drug discovery dataset.

As mentioned earlier in the paper, some of the authors have previously examined the use of "small bags" within the protein secondary structure prediction problem [9]. The small bag approach was compared with creating disjoint partitions of equal size. For the same size partitions, they found that using disjoint partitions was as good or better than the small bagging approach. This paper, in part, extends this work by showing that with a sufficiently large number of small bags subsampled without replication, the performance increases. An important consideration for using this approach as opposed to disjoint partitions is the overall number of subsamples required (the Ensemble Task Size or ETS). Generally the performance of the ensemble does not improve over the baseline C4.5 performance until the ETS value is larger than 1.0. This is equivalent to using a total amount of data greater than the original dataset. However, the key advantage to the subsampling approach is that no single classifier is required to use this amount of data. With an appropriate implementation in a parallel or distributed learning environment, the additional storage overhead is minimal. The ensemble approach is only practical in a distributed environment in which the maximum serial learning task time is the time required to learn over a subsample of the total data or perhaps for very large datasets that cannot fit in a single main memory.

One particular concern within the biological problem domains is the adequacy of a decision tree learner such as C4.5 [8]. Decision trees often consider partitioning of data by choosing a single feature at each split in the tree. In both the representation of a protein chain from amino acids and the structural characteristics of a molecule, we expect the relationships and interdependence of multiple features to be a key descriptive aspect. As a result, more sophisticated algorithms such as neural networks may be more appropriate. For example, a simple feed-forward back-propagation neural network with 75 hidden units produces a per chain accuracy of 72.51% on the Jones test set. The highest accuracy using our subsampling ensemble approach within this paper is 67.67% using 7% subsample sizes. The existence of significant feature dependence is likely the reason for the large improvement gains that are seen in both experiments on the protein prediction dataset. We believe that C4.5 cannot adequately capture these correlations, although the use of ensemble techniques significantly improves the accuracy. Further work on this problem will require classifiers capable of integrating multiple features in the same decision point (i.e. a non axis-parallel decision surface). Many of the classifiers capable of such consideration are extremely expensive in computation time for large datasets and many features. We believe that the use of subsampling is one method of utilizing these more sophisticated algorithms in this large-scale domain.

Over-generalization in learning from the protein structure and drug discovery fields is an important issue. The inherent noise and violation of independence assumptions indicate that over-generalization is a useful technique. We describe a method of using an ensemble of small subsamples as a solution to the over-generalization requirement. Choosing smaller subsamples of data leads to more diverse classifiers, which in combination with the ensemble approach tends to average the hypotheses. Averaging over more potential hypotheses leads to a more general hypothesis overall, with respect to a particular training set. This more general hypothesis is precisely what is needed for the difficult biological domains considered. We demonstrate that this over-generalization technique can lead to classification accuracy gains in both drug discovery and protein secondary structure prediction.

6. ACKNOWLEDGMENTS

This research was partially funded by Tripos, Inc.; the United States Department of Energy through the Sandia National Laboratories LDRD program and ASCI VIEWS Data Discovery Program, contract number DE-AC04-76D000789; and the National Science Foundation, contract number NSF EIA-013-768.

7. REFERENCES

[1] S.F. Altschul, T.L. Madden, A.A. Schaffer, J.H. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, 1997.

[2] Eric Bauer and Ron Kohavi. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Machine Learning*, 36:105–142, 1999.

[3] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.

[4] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[5] L. Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383, 1996.

[6] L. Breiman. Pasting small votes for classification in large databases and on-line. *Machine Learning*, 36:85–103, 1999.

[7] Robert D. Brown and Yvonne C. Martin. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *Journal of Chemical Information and Computer Sciences*, 36(3):572–584, 1996.

[8] N. V. Chawla, T. E. Moore, K. Bowyer, L. O. Hall, C. Springer, and P. Kegelmeyer. Investigation of bagging-like effects and decision trees versus neural nets in protein secondary structure prediction. In *Workshop on Data Mining in Bioinformatics, Knowledge Discovery and Data Mining (KDD)*, 2001.

[9] N.V. Chawla, T.E. Moore, K.W. Bowyer, L.O. Hall, W .P. Kegelmeyer, and C. Springer. Bagging is a small dataset phenomenon. In *Proceedings of the International Conference of Computer Vision and Pattern Recognition (CVPR)*, Hawaii, December 2001.

[10] Jie Cheng, Christos Hatzis, Hisashi Hayashi, Mark A. Krogel, Shinichi Morishita, David Page, and Jun Sese. Kdd cup 2001 report. *SIGKDD Explorations*, 3(2):47–64, January 2002.

[11] Robert D. Clark, David E. Patterson, Farhad Soltanshahi, James F. Blake, and James B. Matthew. Visualizing substructural fingerprints. *Journal of Molecular Graphics and Modelling*, pages 404–411, 2000.

[12] Cynthia Corwin and Irwin D. Kuntz. Database searching: Past, present and future. In Yvonne C. Martin and Peter Willett, editors, *Designing Bioactive Molecules*, pages 1–16. American Chemical Society, Washington, D.C., 1998.

[13] James Darnell, Harvey Lodish, and David Baltimore. *Molecular Cell Biology*. Scientific American Books, 1990.

[14] T. G. Dietterich. Ensemble methods in machine learning. In J. Kittler and F. Roli, editors, *First International Workshop on Multiple Classifier Systems*, Lecture Notes in Computer Science, pages 1–15, New York, 2000. Springer Verlag.

[15] T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, 40(2):139–158, 2000.

[16] Y. Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

[17] Cynthia Gibas and Per Jambeck. *Developing Bioinformatics Computer Skills*. O'Reilly and Associates, 2001.

[18] David T. Jones. Protein secondary structure prediction based on decision-specific scoring matrices. *Journal of Molecular Biology*, 292:195–202, 1999.

[19] E. M. Kleinberg. An overtraining-resistant stochastic modeling method for pattern recognition. *The Annals of Statistics*, 24(6):2319–2349, 1996.

[20] R. Kohavi and D. H. Wolpert. Bias plus variance decomposition for zero-one loss functions. In L. Saitta, editor, *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 275–283. Morgan Kaufmann, 1996.

[21] Lawrence Livermore National Laboratory. Protein structure prediction center. `http://predictioncenter.llnl.gov`.

[22] Louisa Lam and Ching Y. Suen. Application of majority voting to pattern recognition: An analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 1997.

[23] H. Liu and H. Motoda, editors. *Instance Selection and Construction for Data Mining*. Kluwer Academic Publishers, 2001.

[24] Yvonne C. Martin, Peter Willett, et al. Diverse viewpoints on computational aspects of molecular diversity. *Journal of Combinatorial Chemistry*, 3(3):231–250, May/June 2001.

[25] Developmental Therapeutics Program. Diversity set information. National Cancer Institute, National Institutes of Health, `http://dtp.nci.nih.gov`.

[26] Foster Provost, David Jensen, and Tim Oates. Efficient progressive sampling. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pages 23–32, 1999.

[27] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1992.

[28] Peter Willett, John M. Barnard, and Geoffrey M. Downs. Chemical similarity searching. *Journal of Chemical Information and Computer Science*, 38(6):983–996, 1998.