

Disparate Data Fusion for Protein Phosphorylation Prediction

Genetha A. Gray, Pamela J. Williams

Computational Sciences & Mathematics Research Department, Sandia National Laboratories, PO Box 969 MS
9159, Livermore, California 94551-0969, USA, {gagray@sandia.gov, pwillia@sandia.gov}

Kenneth L. Sale

Biosystems Research Department, Sandia National Laboratories, PO Box 969, MS 9292, Livermore, California
94551-0969, USA, {klsale@sandia.gov}

New challenges in knowledge extraction include interpreting and classifying data sets while simultaneously considering related information to confirm results or identify false positives. We discuss a data fusion algorithmic framework targeted at this problem. It includes separate base classifiers for each data type and a fusion method for combining the individual classifiers. The fusion method is an extension of current ensemble classification techniques and has the advantage of allowing data to remain in heterogeneous databases. In this paper, we focus on the applicability of this framework to the protein phosphorylation prediction problem.

Key words: ensemble classification, phosphorylation, base classifier

1 Introduction

Significant advances in methods of data collection coupled with decreasing storage costs have motivated the collection of large volumes of data in areas such as genomics, proteomics, chemistry, and medicine. Extracting meaningful information from these data sets is a challenge that is often hampered by data types that provide different views of the same situation, by data sets that give complimentary information despite appearing dissimilar, and by the wide variety of data collection and storage formats. Our work investigates the use of ensemble classification to address these issues.

Ensemble classification refers to combining the predictions of multiple classifiers into a single classification. Traditionally, these techniques have been used to combine the predictions of different classifiers of the same data set into a single classification. Moreover, this approach has been shown to be more accurate than any of the individual classifiers [1]. In this work, we examine extending the principles of ensemble classification to disparate data sources. Some advantages of this approach are: the data can exist in separate data bases; the data formats do not need to be translated; and there is a savings in computational time and resources.

2 Ensemble Classification of Disparate Data

The theoretical framework of our fusion method is as follows: First, an appropriate data classification scheme is applied to each individual data set. Note that the only restriction placed on the choice of a so-called *base classifiers* is that it be suited to the single data set to which it will be applied. In fact, the appropriate scheme for any one of the data sets may be an ensemble method. Once the classification results are obtained, they are sent to the ensemble or fusion algorithm. It is the job of the fusion algorithm to consider all the base classifications and make an overall global decision about the entire data system. As is the case with the base classifiers, the fusion mechanism could be one of any number of ensemble techniques. The overall goal of this project is to build an underlying framework which would allow users to specify the base classifier and to select an accompanying fusion method.

2.1 Fusion Classifiers

Unweighted majority voting is the simplest and one of the most widely used ensemble classification methods. The algorithm is: Given n data classifications, a decision is accepted if at least k of these classifications agree, where

$$\begin{aligned} k &= n/2 + 1, \text{ if } n \text{ is even,} \\ k &= (n + 1)/2, \text{ if } n \text{ is odd.} \end{aligned}$$

Most classification ensembles are implemented as a means of improving the classification results of a single classifier. In contrast, we use ensembles to combine the results of classifications of disparate, but related data sets. In other words, the unweighted voting algorithm is normally used to describe one data set to which n different classifiers were applied in order to reduce or eliminate classification errors. In the case of disparate data, we consider n individual data classifications that result from n different, but related, sets of data.

Weighted voting [10] is a multiple ensemble classification that is closely related to unweighted voting. The difference in the two methods is that in weighted voting, each of the n individual classifications is assigned a weight, and the weighted votes are considered when making a decision. Currently, there is no conventional method for choosing these weights. Moreover, it has often been shown that weighted voting does not have any significant advantage over unweighted voting. However, this has only been illustrated in the case of applying an ensemble classifier to one data set. Because we are interested in disparate data sources, the use of weights warrants significant research. We opted to use optimization as a means of selecting the weighting scheme in order to provide a mathematically rigorous definition of our fusion technique. For preliminary testing, we choose the weights as be a convex combination (i.e., nonnegative weights summing to one) that minimized the sum of squared misclassification predictions. Specifically, the weights for a problem with n classifications is the vector x that is the solution to the optimization problem:

$$\min \sum_{i=1}^m \left(\sum_{j=1}^n E_{ij} x_j \right)^2 \quad \text{s.t.} \quad \sum_{j=1}^n x_j = 1, \quad x_j \geq 0 \quad \forall j \quad (1)$$

where m is the number of observations and E is a given $m \times n$ matrix. This sort of optimization is easy to do with various existing constrained optimization packages. Thus, the interesting question becomes choosing the matrix E . In our numerical experiments, we first consider a matrix E in which $E_{ij} = 1$ if the j -th classifier computes the i -th observation correctly. Because some classifiers compute a numeric prediction in the course of their decision-making, we also consider matrices which take these predictions into account. This is described in more detail in [8].

Ongoing and future work includes identifying additional appropriate algorithms to be included in the fusion framework.

3 Protein Phosphorylation Prediction

Protein phosphorylation is a biological reaction in which a phosphate group (PO_4) is added to a protein. It is arguably the most important regulatory cellular event. For example, phosphorylation is a key trigger at many stages of immune response pathways. Therefore, it is a key component of understanding immune responses at the cellular level. Prediction of phosphorylation sites is critical to both uncovering the immune response pathway and understanding where in the pathway pathogens may be circumventing the immune response. Data mining and machine learning techniques have become important tools in the task of prediction.

Within a protein, phosphorylation can occur on several amino acids. In this study, we are interested in phosphorylation on only three amino acids: serine, threonine and tyrosine. Serine (S) is the most common type of phosphorylation followed by threonine (T). Tyrosine (Y) is relatively rare but is well understood thanks to the existence of a purifying antibody that simplifies experimental work. Most existing classifiers for S/T/Y identification perform well with ten-fold cross-validation on training sets. However, in practice, they tend to produce many false positives. Thus, our goal in applying fusion classification is to increase the accuracy of phosphorylation prediction for proteins not included in the training sets of the current classifiers.

3.1 Base Classifiers

To begin testing our fusion framework, we selected the following existing phosphorylation prediction methods as base classifiers:

- NetPhos 2.0 server ([2]): uses artificial neural networks trained for each of the three amino acids of interest to predict phosphorylation sites. The serine neural network is trained on an 11-residue window, whereas the threonine and tyrosine neural networks are trained with 9-residue windows. Sites with scores exceeding the threshold value of 0.500 are classified as phosphorylation sites, with values closer to 1.00 indicating a higher confidence that a phosphorylation site has been identified. (www.cbs.dtu.dk/services/NetPhos)
- KinasePhos ([6]): uses hidden Markov Model (HMM) theory to learn phosphorylation sites based on 9-residue windows. This classifier offers a number of options that allow the user to tailor the search. For example, searches can be completed for only serine, threonine, or tyrosines sites or for any combination of the three residues. In addition, the user can define a non-specific kinase prediction or select from a menu of 18 kinases. Finally, the user can select either prediction specificity or HMM bit score for the prediction criterion. (<http://kinasephos.mbc.nctu.edu.tw>)
- Scansite ([11]): finds motifs within proteins likely to be phosphorylated. Unlike the other prediction methods listed, Scansite is not based on a machine learning algorithm. Instead, the package assigns bit scores to each 15-window residue based on position specific scoring matrices whose entries indicate the preference for the amino acid type at each position in the motif. Scansite 2.0 contains 62 position specific scoring matrices where scores of 0.00 indicate an exact match of the motif description. (<http://scansite.mit.edu/>)

Other existing amino acid sequence classifiers could also be used as base classifiers. Some examples include: (i) DisPhos which uses intrinsic disorder information to discriminate between phosphorylation and non-phosphorylation sites, (ii) Predikin which uses similarity rules, (iii) GPS which combines similarity scoring with a clustering algorithm [15], and (iv) PPSP which uses Bayesian decision theory [14].

Although these classifications are useful, they utilize only the linear sequence of amino acids in the proteins. It is well known that three dimensional structure is also important for phosphorylation site prediction. Hence, future work includes augmenting sequence-only base classification methods with information about solvent accessibility at particular sites. Our fusion framework allows combination of these vastly different kinds of information and would allow scientists to make stronger predictions.

Residue	Method	Sensitivity (%)	Specificity(%)	Accuracy(%)
Serine (S)	NetPhos	81.6	53.1	53.8
	KinasePhos	63.3	78.4	78.0
	Scansite	40.8	86.9	85.8
Threonine (T)	NetPhos	70.7	78.2	78.2
	KinasePhos	56.1	87.5	87.2
	Scansite	43.9	89.9	89.5
Tyrosine (Y)	NetPhos	67.4	69.9	69.9
	KinasePhos	53.9	79.4	78.4
	Scansite	44.0	90.2	88.4

Table 1: Comparing the performance behavior of the 3 base classifiers

4 Numerical Results

Our data set contains 1805 protein entries from PhosphoELM version 4.0 ([5]), an open source database of experimentally verified S/T/Y phosphorylation sites. There are 3175 serine, 767 threonine, and 1372 tyrosine known instances of phosphorylation sites in the the set. We follow the standard approach and form our negative set (non-phosphorylation sites) from all the residues centered at S/T/Y that have not been experimentally verified. Consequently, the negative set is orders of magnitude larger than the positive (phosphorylation site) set.

Researchers in the literature use sensitivity (also known as recall), specificity, and accuracy to compare the performance behavior of classifiers on common datasets. Sensitivity denotes the ratio of correctly identified phosphorylation sites to true phosphorylation site and is defined

$$\text{Sensitivity} = \frac{Tp}{Tp + Fn},$$

where Tp is the number of true positives and Fn is the number of false negatives. Specificity is the ratio of correctly identified non-phosphorylation sites to true non-phosphorylation sites and is defined

$$\text{Specificity} = \frac{Tn}{Tn + Fp},$$

where Tn and Fp are the number of true negatives and false positives, respectively. For classes of equal size, accuracy can be computed as the average of sensitivity and specificity. If the class sizes are imbalanced, as they are in our study, a better measure of accuracy is

$$\text{Accuracy} = \frac{Tp + Tn}{Tp + Fp + Tn + Fn}.$$

In PhosphoBase [3], the precursor to PhosphoELM, the ratio of experimentally verified phosphorylation sites to non-phosphorylated sites was 1:31.

For our numerical experiments, we randomly selected 100 proteins from the PhosphoELM database. In addition, we used the default settings in NetPhos, selected 100% specificity as the KinasePhos prediction criterion, and predicted phosphorylation sites with medium stringency in ScanSite. Medium stringency means a match is reported if the score falls within the top one percent

Residue	NetPhos	Scansite	KinasePhos	Weighted Voting I	Weighted Voting II
Serine (S)	3700	1032	1704	1310	723
Threonine (T)	1075	497	617	537	176
Tyrosine (Y)	675	219	462	363	141
S/T/Y	5450	1748	2783	2410	1040

Table 2: Tallies for the number of false positives for the base classifiers and fusion methods

of scores when compared to the vertebrate sequence in the Swiss-Prot protein database [4]. In Table 1, we report the sensitivity, specificity, and accuracy scores for each of the three base classifiers.

Recall that the biggest problem in the prediction of protein phosphorylation sites is the high number of false positives. Therefore, to investigate the usefulness of ensemble classification for this problem, we must first investigate whether or not the voting algorithms can reduce the number of false positives. Table 2 gives a summary of the false positive counts for each of the three base classifiers. It also includes the false positive counts for two weighted voting fusion methods—weighted voting I which gives the worst case scenario of unweighted voting, and weighted voting II which uses the accuracy to obtain the weights. Note that the fusion methods were successful in reducing the overall numbers of false positives. Our hope is that this trend will continue as we incorporate additional data types.

5 Discussion and Future Work

Our goal for our fusion framework is to find and test other methods of ensemble classification appropriate for disparate data fusion. For example, we are currently investigating techniques based on Dempster-Schafer theory that have been successfully applied to sensor fusion problems [9]. Moreover, we hope to use recent work examining the mathematics underlying the complexities of learning from disparate data sources [12] to develop new methods.

In applying our theoretical framework to a real problem, we learned many lessons about the generalities that exist in our framework. For example, two of the three base classifiers output the results in the same format. However, the third used an entirely different format. Therefore, although the fusion methods do not require the actual data sets, some level of standardization of the classifier output must still be accomplished. We plan to add this as an interactive component of our framework.

In terms of the phosphorylation prediction problem, future work includes incorporation three-dimensional structural information (as previously discussed) and of phylogenetic information. It has been shown that the same enzymes in different species can have considerably different amino acid sequences while still providing the cell with the same functionality. Therefore, our goal is to improve prediction of phosphorylation sites by identifying functionally similar proteins in multiple species, making phosphorylation site predictions for each sequence, and then combining these results with measured or predicted structural data. To test this approach, we will make use data from multiple eukaryotes and more than 200 bacterial species described in [7] and [13] to identify proteins and functional modules that remain together across speciation events through evolutionary time.

6 Acknowledgments

The authors would like to acknowledge Joshua Griffin for his development of Ruby scripts to compare the experimentally verified and predicted phosphorylation sites. His efforts saved us countless hours of work.

References

- [1] Banfield, Hall, Bowyer, Bhadoria, Kegelmeyer, and Eschrich. A comparison of ensemble creation techniques. In *Fifth Workshop on Multiple Classifier Systems (MCS 2004)*, pages 223–232, June 2004.
- [2] N. Blom, S. Gammeltoft, and S. Brunak. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Molec. Bio.*, 294(5):1351–1362, 1999.
- [3] N. Blom, A. Kreegipuu, and S. Brunak. PhosphoBase: a database of phosphorylation sites. *Nucleic Acids Research*, 26:382–386, 1998.
- [4] B. Boeckmann and et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Research*, 26:382–386, 1998.
- [5] F. Diella et al. Phosphoelm: A database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, 5:79–83, 2004. <http://phospho.elm.eu.org/>.
- [6] H. D. Huang et al. KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Research*, 33:226–229, 2005. <http://kinasephos.mbc.nctu.edu.tw/>.
- [7] J. M. Stuart et al. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302, Oct. 2003.
- [8] G. Gray, P. Williams, K. Sale, and D. Gay. Enhancing information extraction by applying ensemble classification to disparate data sets. Technical report, Sandia National Labs, Livermore, CA, 2006.
- [9] D. Koks and S. Challa. An introduction to bayesian and dempster-shafer data fusion. Technical Report DSTO-TR-1436, Defence Science and Tech Org, Edinburgh, Australia, Aug 2003.
- [10] N. Littlestone and M. Warmuth. The weighted majority voting algorithm. *Information and Computation*, 108:212–261, 1994.
- [11] J. Obenauer, L. Cantley, and M. Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Research*, 31(13):3635–3641, 2003.
- [12] R. Schuller, S. Ben-David, and J. Gehrke. A theoretical framework for learning from a pool of disparate data sources. In *Proceedings from the 2002 KDD Conference*, pages 443–449, 2002.
- [13] B. Srinivasan. Genome annotation through phylogenomic mapping. *Nat. Biotech.*, 23(6), 2005.
- [14] Y. Xue, A. Li, L. Wang, H. Feng, and X. Yao. PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics*, 7(163), 2006.
- [15] Y. Xue, F. Zhou, M. Zhu, G. Chen, and X. Yao. GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Research*, 33, 2005.