Protein Subcellular Localization Prediction with Associative Classification and Multi-class SVM

Yifeng Liu*, Zhaochen Guo, Xiaodi Ke, Osmar R. Zaïane
Department of Computing Science, University of Alberta, Canada
{yifeng,zhaochen,xke,zaiane}@cs.ualberta.ca

ABSTRACT

Protein subcellular localization prediction is the problem of predicting where a protein functions within a living cell. In this paper, we apply associative classifications (CMAR and CPAR) and multi-class Support Vector Machines to tackle the problem of protein subcellular localization prediction. We use classification feature sources generated from a protein's SwissProt annotation record. We visualize the applied classification rules in an explain graph for domain experts to interpret. We compare the performance of our approaches to those of Proteome Analyst 3.0, using the same set of classification features; we find that all three classification algorithms outperform Proteome Analyst. Multi-class SVM achieves overall F-measures [0.934 $\sim$ 0.991], while CPAR and CMAR achieve overall F-measures [0.922 $\sim$ 0.989] and [0.880 $\sim$ 0.989], respectively. Our result shows that despite multi-class SVM is still the most accurate prediction algorithm with overall F-measures, CPAR and CMAR achieve very similar accuracy. In most cases, CPAR outperforms CMAR, especially when the feature space is large. Our result indicates that associative classification algorithms, especially CPAR, is a good alternative to SVM with similar accuracy but much better transparency in classification models.

Categories and Subject Descriptors

I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems—*Medicine and science*

General Terms

Experimentation, Algorithm, Languages

Keywords

protein annotation, associative classification, support vector machine, text mining, bioinformatics

*To whom correspondence should be addressed

ACM-BCB '11, August 1-3, Chicago, IL, USA

1. INTRODUCTION

Predicting the location (subcell label) where a protein functions within a living cell is essential for elucidating its functions. In the past two decades, researchers tackled the problem of protein subcellular localization prediction using machine learning algorithms with such classification features as N-terminal sorting signals, amino acid composition and textual annotation features [8]. Subcell label prediction is a hard problem due to the scarcity of training data, annotation errors and difficulties inherited from the intricate cell structures. To date, there are on-going efforts in the field of bioinformatics to improve prediction accuracy, refine localization subcategories, and broaden the category of applicable organisms for subcell label prediction.

In recent years, SVM gained popularity as the dominating algorithm for subcell label prediction, thanks to its high accuracy and ability to handle a large collection of classification features [8]. Despite its high accuracy, the trained models and predicted results obtained using SVM are hard to explain due to the cryptic nature of the classification algorithm and the resulting models. *Associative Classification* (AC) is a novel classification approach that combines both *association rule mining* and *supervised classification* [6]. Given labelled training data, an associative classifier finds the features that often co-occur with class labels, and generates classification rules mapping features to class labels using techniques in *association rule mining*. An association rule takes the form of $\{f_1, f_2 \dots f_m\} \rightarrow L$. The left hand side $\{f_1, f_2 \dots f_m\}$ of the rule is called *antecedent*, which is a vector of m features; the right hand side of the rule is called *consequent*, which is a single class label L. Each association rule is a deduction from *antecedent* (features) to *consequent* (class label). The collection of discovered association rules forms a model for supervised classification. These discovered association rules may be pruned to reduce the model size and to increase prediction accuracy. *Support* and *confidence* are two parameters that are often used in rule discovery, rule pruning and prediction. Once trained, an associative classifier is capable of using the discovered rules to predict class labels for unseen test cases. AC is a promising classification approach thanks to its high accuracy, efficiency and transparency.

We show that in this paper, AC is as accurate as SVM, but the resulting models are much more transparent and easier to understand. Our work focuses on applying association classification methods to protein subcellular localization prediction; we also experiment with multi-class SVM with various kernels. Given a query protein, we generate features

Table 1: Statistics for the PA datasets

Organism Type	Class	Instances	Features
Animal	9	15,515	3,861
Plant	9	4,574	1,663
Fungi	9	2,873	2,460
Gram+ Bacteria	3	2,969	1,280
Gram− Bacteria	5	6,168	2,311

from the query protein's SwissProt annotation record [3], then we train associative classifiers and multi-class SVM and compare their performances with Proteome Analyst 3.0 [8], which uses a collection of binary SVM. Finally, we visualize the predictions in an *explain graph* for each AC prediction to help them determine the reliability of a particular prediction.

2. METHODOLOGIES

We evaluate our approach using the publicly available Proteome Analyst datasets [2]. Table 1 shows important statistics for the PA datasets generated with proteins from five different organism types (animal, plant, fungi, gram-positive bacteria and gram-negative bacteria). Similar to Proteome Analyst, we retrieve a query protein's top three homologues using BLAST [4], and extract the *keyword*, *Subcellular Localization* and *InterPro* fields from these homologues' SwissProt annotation record. These three record fields have been selected in Proteome Analyst's previous experiments, and they are reported to be the most relevant fields for subcell label prediction [8]. We purposefully use only these selected fields to ensure our performances are directly comparable to those of PA.

We evaluate all supervised classifiers using stratified 5-fold cross validations for direct comparison with the performance of Proteome Analyst 3.0. We report the classification accuracy as *precision*, *recall* and *F-measure* We experiment with two different types of associative classification algorithms: CMAR (Classification based on Multiple Association Rules) [7] and CPAR (Classification based on Predictive Association Rules) [9]. For associative classification with CMAR and CPAR, we use the LUCS-KDD software library [1] with modifications to perform 5-fold cross validation and to extract classification rules for each prediction. For multi-class SVM, we use the LIBSVM package [5]. We optimize the performance of all classifiers within our limit of computational resources. We run multi-class SVM with *linear*, *polynomial* and *RBF* kernels; we also optimize CMAR by varying *support* and *confidence* parameter, and CPAR by varying *minimum best gain* and *gain similarity ration*.

Finally, we transform the applied rules for each prediction into a graphical representation (called *explain graph*) automatically using an in-house computer program written in the `Python` programming language with the `GraphViz` plotting program. Figure 2 shows such a graph as an example.

3. RESULTS AND DISCUSSIONS

In this section, we present our results and discuss interesting issues. Overall precision, recall and F-measure for each classifier using SwissProt annotation features are shown in Table 2 and Figure 1 with comparison to PA 3.0 [8]. We also experiment with feature generated from a protein's referencing PubMed abstracts and amino acid compositions;

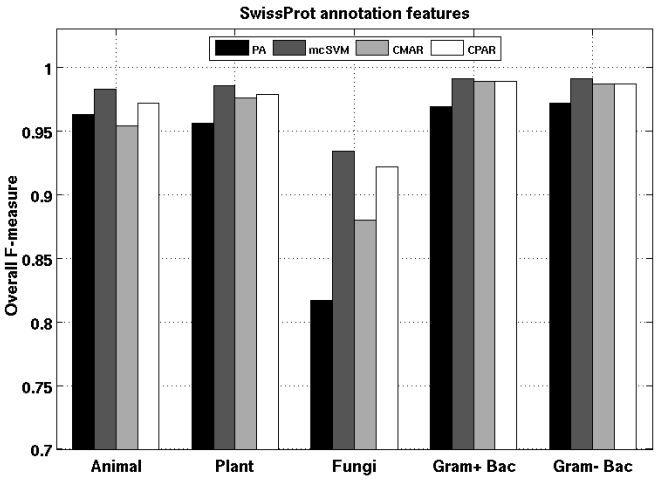


Figure 1: Overall F-measures with various features.

however the results with these additional features are consistently worse than those with SwissProt annotation features and are omitted due to space limitations.

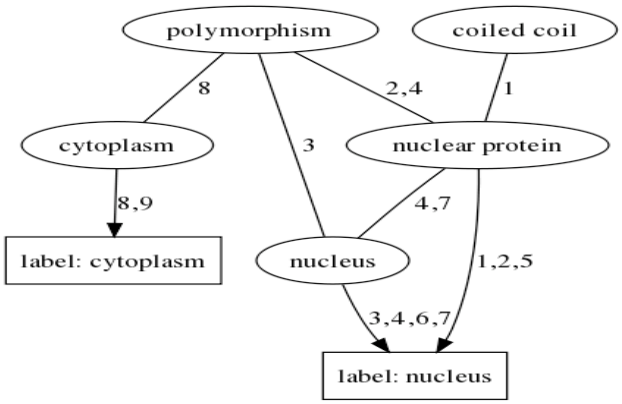
Both CMAR and CPAR outperform Proteome Analyst or at least achieve competitive performance. However, neither CMAR or CPAR significantly outperform PA. Multi-class SVM achieves the best performance for all organism types, but the differences between multi-class SVM and any other predictors are not significant. CMAR and CPAR are both slightly worse than multi-class SVM in Animal, Plant and much worse in Fungi. The performances of CMAR and CPAR are almost indistinguishable, except CPAR outperforms CMAR by roughly 4% in Fungi. The differences between CMAR and CPAR are not significant. In summary, all classification methods in our approach outperform or perform competitively well with Proteome Analyst 3.0 using the same set of features.

While AC and SVM are two totally different types of algorithms, we can still compare them in terms of prediction accuracy, efficiency and transparency. First of all, associative classification is more transparent than SVM, in the sense that AC is capable of showing the classification rules used for each prediction and the strength of each rule; AC could also be modified to allow users to manually edit the trained classification model. On the contrary, SVM can only provide the user a cryptic probability score derived from decision hyperplanes, which are very difficult to visualize. Second, AC is as efficient and accurate as SVM. For example, CPAR only generates a small set of highly selective rules, but it achieves similar accuracy as both binary SVM (as in PA 3.0) and multi-class SVM with SwissProt features. However, a principle drawback of AC is its inefficiency in handling a very large feature space.

Our result shows that multi-class SVM performs universally better than Proteome Analyst 3.0, which is a collection of binary SVM. PA trains a collection of Binary SVM for each organism type, dedicating each binary SVM to each class label. As a result, to predict the most probable subcell label for a single query protein, PA needs to train one model (binary SVM) for each class label. Conversely, with multi-class SVM, we only need to train one model for each organism type (with all class labels), thus saving time in

Table 2: Performance comparison for all classifiers. Best results are shown in bold.

	Precision				Recall				F-measure			
Organism	**PA**	**SVM**	**CMAR**	**CPAR**	**PA**	**SVM**	**CMAR**	**CPAR**	**PA**	**SVM**	**CMAR**	**CPAR**
Animal	0.970	**0.983**	0.963	0.972	0.956	**0.983**	0.945	0.972	0.963	**0.983**	0.954	0.972
Plant	0.968	**0.986**	0.981	0.979	0.945	**0.986**	0.972	0.979	0.956	**0.986**	0.976	0.979
Fugni	0.857	**0.934**	0.913	0.922	0.765	**0.934**	0.849	0.922	0.817	**0.934**	0.880	0.922
Gram+ Bac	0.980	**0.991**	**0.991**	0.989	0.959	**0.991**	0.988	0.989	0.969	**0.991**	0.989	0.989
Gram− Bac	0.984	**0.991**	0.990	0.987	0.960	**0.991**	0.984	0.987	0.972	**0.991**	0.987	0.987




```
Class     : nucleus
Prediction: nucleus

Features  : polymorphism, glycoprotein, coiled coil,
            cytoplasm, nuclear protein, nucleus,
            ipr010978, ipr009053, ipr006933

Classification Rules:

Rule 1: {coiled coil, nuclear protein} -> nucleus (90.45%)
Rule 2: {polymorphism, nuclear protein} -> nucleus (93.97%)
Rule 3: {polymorphism, nucleus} -> nucleus (94.86%)
Rule 4: {polymorphism, nuclear protein, nucleus}
                                        -> nucleus (95.68%)
Rule 5: {nuclear protein} -> nucleus (95.91%)
Rule 6: {nucleus} -> nucleus (96.06%)
Rule 7: {nuclear protein, nucleus} -> nucleus (96.7%)
Rule 8: {polymorphism, cytoplasm} -> cytoplasm (86.2%)
Rule 9: {cytoplasm} -> cytoplasm (89.7%)
```

Figure 2: An explain graph visualizing the applicable rules for an animal protein ith SwissProt annotation features. *Confidence* for each rule is shown in arenthesis.

both training and prediction, while achieving a higher prediction accuracy.

We found that CPAR is more efficient and more accurate than CMAR in most cases. CPAR greedily generates a highly selective set of rules while CMAR generates all rules above the *support* and *confidence* thresholds before filtering them. As a result, we observe that CPAR generates far less rules than CMAR in our experiments, yet achieving higher prediction accuracy.

4. CONCLUSION

In this paper, we apply two associative classification algorithms (CMAR and CPAR) and multi-class SVM for predicting protein subcellular localizations using classification features generated from SwissProt annotation records. Our result shows that both multi-class SVM and CPAR outperform Proteome Analyst 3.0 [8]. Multi-class SVM is still the most accurate classification algorithm; however CPAR is as accurate as multi-class SVM in most cases, and in some cases more robust to noise in the feature sets. CPAR is therefore a good alternative to SVM in protein subcell label prediction. In addition to our effort in optimizing the performance of CMAR and CPAR, we also focus on explaining the predictions of association classifications. We propose a framework of explaining subcell label predictions with associative classification by visualizing the classification rules to help domain expert better interpreting the classification process. We conclude that associative classification has the potential to achieve similar accuracy as SVM, but with much better prediction transparency.

5. ACKNOWLEDGMENTS

We thank the Proteome Analyst Research Group for sharing with us their precious datasets and source code. We also thank the referees for their comments and suggestions.

6. REFERENCES

[1] The LUCS-KDD software library. http://www.csc.liv.ac.uk/~frans/kdd/software/.

[2] The Proteome Analyst 3.0 dataset. http://webdocs.cs.ualberta.ca/~bioinfo/pa/datasets.html.

[3] Uniprot knowledge base. http://www.uniprot.org.

[4] S. F. Altschul, T. L. Madden, A. A. Schaffer, 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 Research*, 25:3389–3402, 1997.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[6] H. Fadi Thabta. A review of associative classification mining. *The Knowledge Engineering Review*, 22(01):37–65, 2007.

[7] W. Li, J. Han, and J. Pei. CMAR: accurate and efficient classification based on multiple class-association rules. *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 369–376, 2001.

[8] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. S. Wishart, B. Poulin, C. Macdonell, and R. Eisner. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, 20(4):547–556, 2004.

[9] X. Yin and J. Han. CPAR: Classification based on predictive association rules. *SIAM International Conference on Data Mining (SDM'03)*, 2003.