International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-3, Issue-5)

Research Article

May
2014

A Novel Data Mining Approach for Protein Function Prediction

Ankita Srivastava[1] & Yogesh Kumar[2]
Department of Computer Science and Engineering
Galgotias University, Greater Noida , U.P., India [1, 2]

Abstract:

Protein function prediction is very important and challenging task in Bioinformatics. In this paper we have used proteins represented by a set of enzymes i.e. Oxidoreductase, Transferase, Hydrolase, Isomerase, Ligase, and Lyase, extracted from the Enzyme Commission (EC) classification to build the models. We have used SVM (Support Vector Machine) to predict protein function which is more efficient for resolving classification problems. We have used protein dataset available at PDB using features such as primary structure, molecular weight, structural molecular weight, chain length, atom count, ligand molecular weight and residue count as training parameters and EC number as corresponding output. Here we use expert model of SVM, with RBF kernel function where width is 0.10 and parameter C is 10. The result in this paper using these parameters shows that the overall average accuracy is 88.49%.

Keywords: Proteins, Function Prediction, SVM, Classification.

1. INTRODUCTION

Proteins are formed from a set of 20 amino acids and the function of a protein is closely related to the structure. There are various function of protein such as catalysis, transport and information. Enzyme behaves like a catalyst which speed up the rate of reaction without becoming the part of reaction. The primary structure of a protein is the sequence of amino acids, secondary structure is the formation of alpha helixes, beta sheets and loops and the tertiary structure is responsible for the spatial arrangement of the protein and the quaternary structure refers to the proteins that have more than one chain of amino acids. In this paper we used the proteins that are classified according to EC number. Finding protein function is an important task which supports the research for novel drug design. In this paper we used six classes of enzymes Oxidoreductase, Transferase, Hydrolase, Isomerase, Ligase, and Lyase. In this paper we used seven features primary structures, molecular weight, structural molecular weight, chain length, atom count, ligand molecular weight and residue count to predict the protein function. Using these features we construct the expert model of support vector machine to predict the protein function. Here we describe the previous research work carried out for the protein function prediction or classification and we also discuss about the various classifiers models that are used in this study. L.Y. Han et al.. [1] proposed a method to predict functional family of protein that is useful for protein function prediction. Every protein sequence is represented by a set of amino acid composition by using these composition he used SVM, supervised machine learning and the result of this model is compared with the Naïve Bayes and C4.5. C.Z. Cai et al. [3] used the SVM for protein function classification. He used a various protein classes such as RNA-binding, homodimer, drug absorption, drug excretion etc. He found the testing accuracy between 84-96%. Paul D. Dobson et al. [2] proposed a method that can assign the function from the structure of protein by using EC number. He used one-class versus one-class SVM to predict the protein function. He found the accuracy between 35-60%. Luiz C. Borro et al. [4] proposed a method for predicting EC number. He used various features of the protein structure find from STRING_DB and used Bayesian classifier to predict the protein function. He found the accuracy 45.3%. Yong-Cui Wang et al. [5] proposed a method to predict enzyme functions using amino acid composition, their neighborhood relationship to each other, and the hierarchical structure of the class. He compared the results from the attributes considered and concludes that the information from all three together offers better results. Using the SVM classifier, they obtain a prediction rate of between 81% and 98%.

2. METHO DOLOGY

2.1 SUPPORT VECTOR MACHINE

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Support Vector Machine (SVM) is a robust classification and regression technique that maximizes the predictive accuracy of a model without over-fitting the training data. SVM works by mapping data to a high dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, and then the data are transformed in such a way that the separator could be drawn as a hyper plane. Following this, characteristics of new data can be used to predict the group to which a new record should belong. After the transformation, the boundary between the two categories can be defined by a hyper plane. The mathematical function used for the transformation is known as the kernel function. SVM supports the Linear, Polynomial, Radial basis function (RBF) and Sigmoid kernel types. When there is a straightforward linear separation of then linear function is used otherwise we used Polynomial, Radial basis function (RBF) and Sigmoid kernel function. Besides the separating line between the categories, a classification SVM model also finds marginal lines that define the space between the two categories. The data points that lie on the margins are known as the support vectors. The margin will be wider between the two categories for better model and prediction of new record. When the margin is not wider than the model is called over-fitted. The parameters of the SVM are highly sensitive and vary for each problem and data sets.

3. DATA SETS

The protein raw data set used in this paper is obtained from PDB. We had taken 4000 data set of protein enzymes from PDB classified according to EC Number and Enzyme name and after that we reduce the reductant data set by selecting the sequence that are similar and after that we remove that data sets whose chain length is less than 100. Seven features, primary structure (sequences), molecular weight (MW), ligand molecular weight, structure molecular weight, residue count, atom count, chain length are extracted from PDB. Table 1 shows the description of the data set and Table 2 shows proteins according to class and total data sets taken for training and testing. Data preparation and all manipulation have been done using Microsoft Excel.

Table 1: Data set description

S.NO	Fields	Description
1	Sequence	The linear amino acid sequence of a protein
2	Molecular weight	Protein molecular weight
3	Ligand molecular weight	Ligand (Ion or Functional molecule) molecular weight
4	Structure molecular weight	Structure molecular weight
5	Residue Count	Residue count of protein
6	Chain Length	Polymer chain length
7	Atom Count	Atom count of protein

Table 2: Data set description of six enzymes

EC.NO	Class(Enzymes)	Function	Total Set
1	Oxidoreductases	Catalyze the reduction oxidation reactions.	275
2	Transferases	Transfer a functional grouping and a donor group to a receptor	801
3	Hydrolases	Hydrolases Catalyze hydrolysis, the breaking of links and structures by the action of water.	627
4	Lyases	Enzymes which catalyze the cleavage of C-C, C-O and C-N links.	316
5	Isomerases	Catalyze the isomerization reactions of simple molecules.	220
6	Ligases	Formation of links by condensation of substances.	171

4. Performance Evaluation

The performance of support vector machine is measured by the quantity of True positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). Where TP (True Positive) is the number of positive instances that are classified as positive, FP (False Positive) is the number of Negative instances that are classified as positive, TN(True Negative) is the number of Negative instances that are classified as Negative and FN(False Negative) is the number of positive instances that are classified as Negative. By using these quantities standard Accuracy Sensitivity, Specificity and Precision performance measure is defined as follows.

Accuracy= (TP+TN)/ (P+N) -- The proportion of instances that are correctly classified.
Sensitivity=TP/P--The proportion of positive instances that are correctly classified as positive.
Specificity=TN/N--The proportion of negative instances that are correctly classified as negative.
Precision=TP/ (TP+FP)--The proportion of instances classified as positive that are really positive.

It is described as the following table

	Predicted class		
	Positive	**Negative**	**Total**
Positive	**TP**	**FN**	**P**
Negative	**FP**	**TN**	**N**

5. RESULT AND DISCUSSION

For raw data manipulation Microsoft Excel is used. We implemented the SVM using SPSS Clementine 12.0 computing environment. Firstly the model is trained with training data set and then tested by testing and validation datasets. In this paper we use 5 fold cross validation method to measure the performance of the support vector machine.

(a) Result Obtained for the six classes of protein analyzed

Class	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
Oxidoreductases (EC-1)	91.63	83.27	100	100
Transferases (EC-2)	98.13	96.26	100	100
Hydrolases (EC-3)	91.56	83.12	100	100
Lyases (EC-4)	84.38	68.77	100	100
Isomerases (EC-5)	68.33	36.65	100	100
Ligases(EC-6)	60.75	21.51	100	100

(b) The Overall Performance Evaluation of Six Classes with 5 fold cross validation.

K FOLD CROSS VALIDATION	**TP**	**FN**	**TOTAOL**	**FP**	**TN**	**TOTAL**
K=1	380	119	499	0	453	453
K=2	357	96	453	0	478	478
K=3	362	116	478	0	480	480
K=4	374	106	480	0	506	506
K=5	387	119	506	0	499	499
TOTAL	1860	556	2416	0	2416	2416

Average Performance Evaluation of Six Classes:

	Positive	**Negative**	**Total**
Positive	1860	556	2416
Negative	0	2416	2416

Accuracy	**Sensitivity**	**Specificity**	**Precision**
88.49	76.99	100	100

Here we observed that the overall accuracy of support vector machine is 88.49%.

6. CONCLUSION

In this paper we proposed support vector machine based method for classification of Enzymes. The result shows that it is capable for classification of different enzymes functions. Here we found the overall accuracy of support vector machine is 88.49%. In this paper we use seven features for the prediction of protein function; in future we can use more features and different features selection algorithms for the prediction of protein function and obtain better result.

REFERENCES

1. L.Y. Han, C.Z. Cai, Z.L. Ji, Z.W. Cao, J. Cui, and Y.Z. Chen, "Predicting functional family of novel enzymes irrespective of sequence similarity," Nucleic Acids Research, vol. 32, pp.6437- 6444, 2004.
2. Paul D. Dobson and Andrew J. Doig, "Predicting Enzyme Class from Protein Structure without Alignments," Journal of JMB, vol. 345, pp. 187-199, 2005.
3. Lu, L., Qian, Z., Cai, Y. D., Li, Y. ECS, "An automatic enzyme classifier based on functional domain composition", Journal of Computer Biol Chem, 31 (3), pp. 226-232, 2007
4. Luiz C. Borro, Stanley R.M. Oliveira, Michel E.B. Yamagishi, AdaultoL. Mancini, Jose G. Jardine, Ivan Mazoni, Edgard H. dos Santos,Roberto H. Higa, Paula R. Kuser, and GoranNeshich, "Predictingenzyme class from protein structure using Bayesian classification," Journal of Genetics and Molecular Research, vol. 5, pp. 193- 202, 2006.
5. Yong-Cui Wang, Yong Wang, Zhi-Xia Yang, Nai-Yang Deng, "Support vector machine prediction of enzyme function with conjoint triad feature and hierarchical context, Journal of BMC Systems Biology, vol. 5, 2011.
6. C.Z. Cai, W.L. Wang, L.Z. Sun, and Y.Z. Chen, "Protein function classification via support vector machine approach," Journal of Mathematical Biosciences, vol. 185, pp. 111-122, 2003.
7. Burbidge, R., Buxton, B., "An introduction to support vector machines for data mining" Journal of operational Research Society: Operational Research Society, March (2001).
8. Wang, Y. C., Wang, X. B., Yang, Z. X., Deng, N. Y.. Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature. *Protein Pept. Lett*., 17, pp. 1441–1449, 2010
9. Qiu, J. D., Huang, J. H., Shi, S. P., Liang, R. P., "An Approach with Support Vector Machine Based on Discrete Wavelet Transform", Journal of Protein Pept Letter, vol. 17, no. 6, pp. 715-22, 2010.
10. Cortes, C. Vapnik, V., "Support-vector networks. Machine Learning", Vol. 3, no. 20, pp. 273–297, 1995