PREDICTION OF DISORDERED REGIONS OF PROTEINS WITH NEW N-PIECES NAÏVE BAYES ALGORITHM

UMUT ORHAN
Computer Center, Gaziosmanpasa University, Tokat, TURKIYE
umutorhan@gop.edu.tr

Irem ERSÖZ
Technical Education Dept, Mersin University, Tarsus, TURKIYE
iremer@mersin.edu.tr

TURGAY IBRIKCI
Computer Engineering Dept, Cukurova University, Adana, TURKIYE
ibrikci@cukurova.edu.tr

ABSTRACT
In this paper, Bayesian Learning with bioinformatics is utilized. Data is partitioned into a different numbers of pieces for Naïve Learning, which is based on binary Naïve Bayes algorithm. The algorithm is applied on the dataset R80 which consists of order/disorder protein structures. In our algorithm, the data is heuristically partitioned into 2, 5, 11, 101, 133, 201, 301, and undivided. The experimental results have indicated that the partition of the data into 133 pieces has given better results according to correct classification and specificity.

INTRODUCTION

The protein macromolecule represents much of the bulk of an organism and accomplishes almost all of its biochemical activities. It has been widely believed that the 3D structure of a protein is a prerequisite for its function. However, the last 50 years of research has revealed that some proteins or protein regions that lack a 3-D structure in the intrinsic state are also involved in a variety of important biological functions [1-4]. These proteins are now called natively unfolded or intrinsically disordered proteins [5]. The research has shown the importance and advantages of disordered proteins for many functional activities such as cell cycle regulation, molecular recognition, enzyme catalysis. Numerous tools and approaches for recognizing and predicting disordered regions of proteins have been developed [6-9]. Identification of these disordered proteins or regions can be done by several experimental techniques such as X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR) but the methods require skilled experts to operate the equipment. This can be quite expensive and time consuming. Thus research has concentrated on study of computational methods including ANNs [10, 11], SVMs [12, 13], logistic regression [14]. Surveys have demonstrated that amino acid composition and other attributes such as hydropathy, flexibility, sequence complexity, net charge, amino acid volumes, and secondary structure can be used as a discriminators for disorder/order propensity [15].

The application of machine learning techniques in bioinformatics problems has become increasingly known in recent years. There is a plethora of intelligent algorithms which are used in various fields. However, many of these algorithms

can be used for only a restricted field. Neural networks and probabilistic prediction algorithms are common functions which can be implemented into a various fields. Bayesian networks are the subset of probabilistic prediction.

Bayesian networks are a useful tool for statistical modeling. They have been widely used in the biological sciences for the tasks of inferring cellular networks [16], modeling protein-signaling pathways [17], data integration, classification, and genetic data analysis [18]. Bayesian networks provide a neat compact representation for expressing joint probability distributions and for inference. The representation and use of probability theory makes Bayesian networks suitable for learning from incomplete datasets, expressing causal relationships, combining domain knowledge and data, and avoid over-fitting a model to training data.

A Bayesian network includes a classifier to show relationship among features. The classifier is a function f that maps the input feature vectors $x \in X$ to the output class labels $y \in \{1, \ldots, C\}$, where X is the feature space. It has been typically assumed that $X = R^D$ or $X = \{0, 1\}^D$, i.e. that the feature vector is a vector of D real numbers or D binary bits, but in general, discrete and continuous features have also been included. The class labels were taken as unordered and mutually exclusive in this study. If an input belongs to multiple classes, this is called a multi-label problem. The goal is to learn f from a labeled training set of N input-output pairs , (x_n, y_n) $n = 1:N$ which has been known as supervised learning [19].

The probabilistic classifiers that recall $p(x \mid y)$ are investigated. The advantages of obtaining probabilities are discussed below. There has been two main ways to attain probability. The first is to learn directly the function that computes the class posterior $p(x \mid y)$, named a discriminative model, since it discriminates between different classes given the input. The alternative is to learn the class-conditional density $p(x \mid y)$ for each value of y and to estimate the class priors $p(y)$; therefore, the Bayesian Rule has been applied to compute the posterior,

$$p(x \mid y) = \frac{p(x, y)}{p(x)} = \frac{p(x \mid y)p(y)}{\sum_{y'=1}^{C} p(x \mid y')p(y')} \tag{1}$$

Naive Bayes is the simplest form of Bayesian network, in which all are independent given value of class variable. This is called conditional independence [19, 22]. Naïve Bayes classifiers currently appear to be particularly popular in commercial and open-source spam filters. This is probably due to their simplicity, which makes them easy to implement, their linear computational complexity, and their accuracy, which in spam filtering is comparable to that of more elaborate learning algorithms [23]. The paper has proposed the N-pieces Naïve for constructing Bayesian networks from prior knowledge. This algorithm has been repeated for different N values (2, 5, 11, 101, 133, 201, 301 pieces and undivided data) in order for validation.

DATASET

In this study, only one dataset named R80 for training and validating processes is used. The dataset R80 including 80 protein taken from the study of Yang et al.[6] has been recreated via the PSSMP method by Su et al. [7].The details of the dataset are given at the Table 1. Each protein chain of the data

contains a region of at least 21 consecutive disordered residues and all were collected from the PDB database. The method has been implemented on the R80 by conserving it without changes to compare the evaluations and measure the performance of Bayesian Learning on protein disorder prediction against the others.

Table 1: Summary of the dataset employed in the study

	Training - Testing Data
	R80
Number of chains	80
Number of ordered regions	151
Number of disordered regions	183
Number of residues in the ordered regions	29909
Number of residues in the disordered regions	3649
Total residues in the dataset	33558

VALIDATION

Predicting a residue in a given protein sequence as order or disorder is a binary classification problem, and many measures have been introduced for validation issues [20, 21]. Table 2 lists five widely used indices defined by previous related works [2-15, 20, 21]. The most important value is the correct classification in this paper. Also mistaken results have also been considered. Sensitivity represents the fraction of disordered residues correctly identified in a prediction method, while specificity indicates the fraction of ordered residues correctly identified. The Matthews' correlation coefficient is a popular measure in many bioinformatics problems [21]. However, sensitivity, specificity, and the Matthews' correlation coefficient are seriously affected by the relative frequency of the target class. Therefore, the above three measures are not suitable for evaluating the performance in isolation. The probability excess is independent of the relative class frequency, and this measure can be reduced to sensitivity + specificity - 1 concisely [6, 7]. In this paper, since these five measures have the same tendency with each other, we are adopted specificity and correct classification. The other methods, however, are shown for comparison purposes.

Table 2: The definition of measures employed in the study

Measure	Abbreviation	Equation
Sensitivity	Sens.	$\frac{TP}{TP+FN}$
Specificity	Spec.	$\frac{TN}{TN+FP}$
Matthew's correlation coefficient	MCC	$\frac{TP\times TN - FP\times FN}{\sqrt{(TP+FN)\times(TN+FN)\times(TP+FN)\times(TN+FP)}}$
Probability excess	Prob. Excess	$\frac{TP\times TN - FP\times FN}{(TP+FN)\times(TN+FP)}$
Correct classification	Corr. Class.	$\frac{TP+TN}{TP+TN+FP+FN}$

The definition of the abbreviations used: *TP* is the number of correctly classified disordered residues; *FP* is the number of ordered residues incorrectly classified as disordered; *TN* is the number of correctly classified ordered residues; and *FN* is the number of disordered residues incorrectly classified as ordered.

N-PIECES NAÏVE BAYES ALGORITHM AND ITS APPLICATION

In fact, this is a simple idea that for digitalization, data is put into N number parts. The algorithm includes two phrases, which are partition of data and to apply classic Naive Bayes Learning classifier. The partition of data is based on heuristical values for N number pieces. Each piece of data is normalized and balanced itself. Data into two pieces is called binary data. These are then put into N balanced part. Thus, we obtain equal weighted N part are obtained. In second phase, the classic Naïve Bayesian Classifier method is applied to each piece of data.

This process was tried on the some different N values (2, 5, 11, 101, 133, 201 and 301). For each partition, the first step is to find the average error for digitalization. The results are shown in Figure 1 that shows relation between the average error and number of N pieces. When N is equal to be $6.3x10^6$, the average error is zero as the raw data. Different pieces data applied with the algorithm to follow their performances of correct classification to compare. This figure is used for finding the optimum N value with Figure 2.

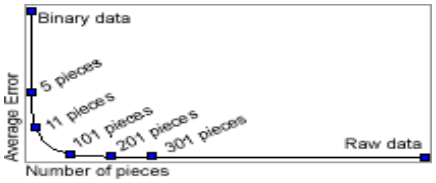

Figure 1: Relation between Average Error and N pieces values

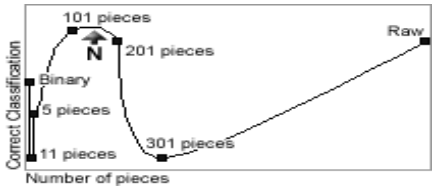

Figure 2: Relation between Correct Classification and N pieces values

When N is equal to 101, the results are promising (Figure 2). But these are not optimum values. The optimum N-pieces was looked sought for the dataset (Figure 3). From Figure 2, it appears that the optimum N value is between 101 and 201, and it is reasonable to use these values. Hence a better point can be obtained as shown by Figure 3.

The optimum N value cannot be 201, since its some other validation measurements are 0. Even 201 value for the correct classification (corr. class.) is better value then others.

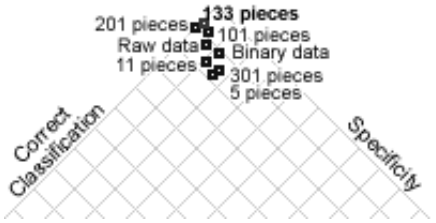

Figure 3: Relation between Correct Classification and Specificity

A study of gradient of line in Figure 2 is useful for the determination of the optimum N value. While the gradient is greater than a threshold, the optimum N value is sought. In Figure 2, the gradient between 201 and 301 is nearly equal to the gradient between 301 and $6.3x10^6$. Hence, the optimum N value cannot be in these regions. It may be between 101 and 201, but what can be an acceptable threshold value? The answer to this question is hidden in Figure 2. The N shows the optimal value which is 133 for the dataset.

Table 3: The performance of each N values on dataset

	Sens.	**Spec.**	**MCC**	**Prob. Excess**	**Corr. Class.**
Binary	0,61	0,91	0,5	0,52	0,87
5 Pieces	0,61	0,87	0,4	0,48	0,84
11 Pieces	0,57	0,90	0,4	0,47	0,86
101 Pieces	0,31	0,97	0,4	0,28	0,90
133 Pieces	0,09	1	0,08	0,09	0,89
201 Pieces	0	1	0	0	0,89
301 Pieces	0,61	0,87	0,4	0,48	0,84
Raw	0,45	0,95	0,4	0,40	0,89

CONCLUSIONS

This paper presented a new Naive Bayesian Learning algorithm which is called NP-NBL. The algorithm is tested on bioinformatics data with good results. The algorithm is applied on the prediction of disorder region of proteins. Many proteins have disorder structures that affect their functional activities. That is why the prediction of disorder structures of proteins are important.

The protein data is partitioned into 2, 5, 11, 101, 201, 301, and raw data. Actually, when we calculate the possibility of partitions which becomes $6.3x10^6$ pieces. According to Figure 2 that shows the critical values are between 101 thru 201. Then we focused on the gap to find the optimal partition of the data. The number of 133 for partition is optimal value with the specificity and the correct classification. However, we obtain the understanding that the optimal N value depending on individual applications.

ACKNOWLEDGMENT

This research was supported by TUBITAK-TBAG-104T505.

REFERENCES

[1] C. Klee, G. Draetta, M. Hubbard. Calcineurin. In Meister, A. (ed.) Advances in Enzymology, Vol. 61, 149–200, 1988.

[2] A.K. Dunker, Z. Obradovic, P. Romero, C. Kissinger, E. Villafranca. On the Importance of Being Disordered. PDB Newsletter, 81, 3-5, 1997.

[3] P.E. WRIGHT and H.J. DYSON. Intrinsically Unstructured Proteins: Re-assessing The Protein Structure-Function Paradigm. J. Mol. Biol., 293, 321-331, 1999.

[4] A.K. Dunker, P. Romero, Z. Obradovic, E.C. Garner, C.J. Brown. Intrinsic Protein Disorder in Complete Genomes. Genome Informatics, 11, 161-171, 2000

[5] A.K. Dunker, J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C.S. Oldfield, A.M. Campen, C.M. Ratliff, K.W. Hipps, J. Ausio, M.S. Nissen, R. Reeves, C. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, W. Chiu, E.C. Garner, Z. Obradovic. Intrinsically Disordered Protein. Journal of Molecular Graphics and Modeling, 19, 26-59, 2001.

[6] Z. R. Yang, R. Thomson, P. McNeil and R. M. Esnouf. RONN: The Bio-Basis Function Neural Network Technique Applied To The Detection of Natively Disordered. Bioinformatics, 21, 3369-3376, 2005

[7] C.T. Su, C.Y. Chen and Y.Y. Ou. Protein Disorder Prediction by Condensed PSSM Considering Propensity for Order or Disorder. BMC Bioinformatics, 7:319, 2006.

[8] J.J.Ward, L.J. McGuffin, K. Bryson, B.F. Buxton, D.T. Jones. The Disopred Server for the Prediction of Protein Disorder. Bioinformatics, 20, 2138-2139, 2004.

[9] P. Romero, Z. Obradovic, A.K. Dunker. Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family. Genome Informatics, 8, 110-124, 1997.

[10] P. Romero, Z. Obradovic, C.R. Kissinger, E. Villafranca, A.K. Dunker. Identifying Disordered Regions in Proteins from Amino Asid Sequence. In Proc. IEEE Int. Conf. On Neural Networks, 1, 90-95, 1997

[11] D.T. Jones, J.J. Ward. Prediction of Disordered Regions in Proteins from Position Specific Scoring Matrices. Proteins, 53, 573-578, 2003.

[12] X. Li, P. Romero, M. Rani, A.K. Dunker, Z. Obradovic. Predicting Protein Disorder for N-, C- and Internal Regions. Genome Inform Ser Workshop Genome Infor., 10:30-40, 1999.

[13] K. Shimizu, S. Hirose, T. Noguchi, Y. Muraoka. Predicting The Protein Disordered Region Using Modified Position Specific Scoring Matrix. Genome Informatics, P150, 2004.

[14] S. Vucetic S, C.J. Brown, A.K. Dunker, Z. Obradovic. Flavors of Protein Disorder. Proteins, 52:573-584, 2003.

[15] P. Romero, Z. Obradovic, X. Li, E.C. Garner, C.J. Brown, A.K. Dunker. Sequence Complexity of Disordered Protein. Proteins: Structure. Function and Genetics, Vol. 42, 38-48, 2001.

[16] N. Friedman. Inferring Cellular Networks Using Probabilistic Graphical Models. 303, 709-805, 2004.

[17] K. Sach, O. Perez, D. Pe'er, D.A. Lauffenburger, and G.P. Nolan. Causal Protein Signaling Networks Derived from Multiparameter Singlecell Data. Science, 308, 523-529, 2005.

[18] M.A. Beaumont and B. Rannala. The Bayesian Revolution in Genetics. Nature Reviews Genetics, 5, 251-261.

[19] K. P. Murphy. Naive Bayes Classifiers. Technical Report, October 2006.

[20] E. Melamud and J. Moult. Evaluation of Disorder Predictions in CASP5. Proteins, 53:561-565, 2003.

[21] Y. Jin and R.L. Dunbrack. Assessment of Disorder Predictions in CASP6. Proteins, Early View, 2005.

[22] V. Metsis, I. Androutsopoulos, G. Paliouras. Spam Filtering with Naïve Bayes–Which Naïve Bayes. CEAS 2006-Third Conference on Email and Anti-Spam, July27-28, California USA, 2006.

[23] I. Androutsopoulos, G. Paliouras, and E. Michelakis. Learning to filter unsolicited commercial e-mail. Technical Report 2004/2, NCSR Demokritos, 2004.