ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 13, October 2016

Greedy Supervised Feature Selection (GSFS) of Physicochemical Properties of Amino Acids

Sagnik Banerjee [1], Angan Mitra[1], Subhadip Basu[1], Mita Nasipuri[1]

Department of Computer Science & Engineering, Jadavpur University, Kolkata, India[1]

ABSTRACT: Feature selection is an immensely important task in the domain of pattern recognition. Physicochemical properties of amino acids have been extensively used in classification and regression tasks like protein secondary structure prediction, protein post translational modification etc. Out of the 544 features, presented in *AAINDEX*, not all are relevant for a certain classification. Classifiers like Neural Network (NN) or Support Vector Machine (SVM) will include information from all the features in order to model a problem. Hence, presenting the most appropriate features to a classifier will enhance its performance. Feature selection algorithms attempt to select a subset of the available features which will help perform classification better. Feature selection methods can be supervised or unsupervised. In this paper we present a greedy supervised algorithm (GSFS) to select the most apt features for a particular classification task. We have implemented our algorithm on the problem of protein phosphorylation of human proteins and have obtained better results than other works which have implemented an unsupervised algorithm for the same task. We have achieved an increase of area under roc curve of about 1% in 5 fold cross validation experiments.

KEYWORDS: Greedy supervised feature selection, GSFS, Phosphorylation, Human proteins, Cross validation, Support Vector Machines.

I. INTRODUCTION

The word coined as 'feature' represents a prominent attribute or aspect of an object, both living and non-living. The presence of these features gives an object its properties to interact and behave. While analyzing a particular aspect the contribution of all these features may not be significant, even selection of some may lead to an erroneous result. Hence the method of selecting which of the features to use is of utmost importance. Definition of feature selection involves taking a subset of characteristics of an object into consideration for a model construction. The prime target of the feature selection is to eliminate the redundant information and thereby reduce the search space.

Subset selection is used to identify the most important columns if all the features can be visualized in a matrix and thus subset selection approach chooses a subset of features and uses them for evaluation as a group for suitability. Subset selection can be formulated as a search algorithm to traverse through the space of possible features and thus evaluate each subset by running a model on the subset. Another approach states that instead of evaluating a selected subset against a model, a simpler filter is evaluated. Greedy Hill Climbing provides another solution in which the algorithm iteratively evaluates a candidate subset of features, modifies the subset and accepts the solution if the new subset is better than the old.

II. RELATED WORKS

A lot of work has been previously done on feature selection. Prior to classification, feature selection is applied to choose a subset of features to enhance the quality of the prediction model. Therefore, feature selection has found its application in computational biology, image processing, optical character recognition, facial emotion recognition etc. Saha et. al has performed fuzzy clustering of the physicochemical and biochemical properties of the amino acids mentioned in *AAINDEX* in [1]. They have adopted an unsupervised scheme and have created 3 groups of selected features each containing 8, 24, 40 features respectively. As a result of the clustering eight clusters were formed. The centroids of each cluster were chosen as the first subset resulting in 8 features. Next they suggested another group where, from each cluster two more features were incorporated which were farthest in the corresponding cluster. This

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 13, October 2016

led to the second group which had 24 features. Finally, in the last group two more features were considered from each cluster which were closest to their corresponding cluster centre. Alexandros Kalousis, Julien Prados, Melanie Hilario in [2] have performed a series of experiments with several feature selection algorithms on a set of proteomics datasets which enabled them to explore the merits of each stability measure and create stability profiles of the feature selection algorithms. A novel approach has been laid down in [3] which makes multiple runs of SVM-RFE (Recursive Feature Elimination) with different thresholds. Each run produces one feature subset. The resulting feature subsets are combined in order to obtain a robust result/classification.

Normalized mutual information feature selection (NMIFS) has been explored in [4] to classify breast cancer as malignant or benign. They have implemented SVM-based recursive feature elimination (SVM-RFE) with NMIFS filter (SRN). With 10 fold cross validation, SRN performs better than other feature selection algorithms. Class-dependent density-based feature elimination (CDFE), for binary data sets have been addressed in [5]. They have combined feature subset selection with CDFE and have successfully reduced the feature set by 97. Feature selection has been applied in research works like post translational modification [6] [7]. The algorithm can be applied on several pattern recognition tasks like protein secondary structure prediction [8], protein-protein interaction [9][10] etc.

It is evident that there does not exist a single feature selection algorithm which performs competitively well in all situations. This paper attempts to address the issue of selecting the best features which would be beneficial for proteomics.

III. BACKGROUND

Pattern recognition is the task of detecting regularities in data. One of the most important task in pattern recognition is classification. Classification is basically a task of mapping an n dimensional vector $X=\{x_1,x_2,\ldots\ldots,x_n\}$ to one of the m points in $Y=\{0,1,2,\ldots\ldots,m\}$. Classifiers like Neural Network, SVM, Naïve Bayes etc. will classify data based on all the supplied features. Sometimes, all the features are not required to correctly classify a data sample. For example, while assessing the academic success of an individual it is immaterial to consider their height and weight. But height and weight are one of the most important features when the health of an individual is considered. Hence, for different classification problems involving the same individual, different sets of features are relevant. Therefore, feature selection algorithms aim to choose a subset of features which will help classify the data with maximum accuracy. It is essential to remove irrelevant features because they often lead to increased computational cost. Also irrelevant features may lead to over-fitting.

A. PHYSICOCHEMICAL PROPERTIES OF AMINO ACIDS

Each of the 20 amino acids possesses some properties like hydrophobicity, molecular size, polarity etc. these properties have been quantified in *AAINDEX* leading to a total of 544 indices. In this paper we have experimented on phosphorylation of human proteins with the help of feature selection. Only a subset of these physiochemical properties are related in obtaining the end result while the others bear no or little semblance with the context of phosphorylation.

B. SUPPORT VECTOR MACHINE

In machine learning, SVM [11] are supervised learning models to analyze data and recognize patterns to be used in regression analysis and classification. Training an SVM is a method of supervised learning and is provided with a suite of training data where there is an input dataset and its corresponding output. The learning algorithm uses this dataset to generalize mappings for it to predict on unseen data. These SVM have a set of algorithm for pattern analysis which are known as kernel. There are 6 types of kernel which are Linear Kernel, Fisher Kernel, Graph Kernel, Polynomial Kernel, RBF Kernel and String Kernel. The class of kernel methods gives an essence of exploring possible patterns by looking into the similarity of the data.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 13, October 2016

C. CROSS VALIDATION

Cross validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly employed to find out the best training parameters for a certain machine like NN or SVM. There exists no mathematical approach to determine the training parameters for a machine making cross validation indispensable. In *k*-fold validation, the entire data is partitioned into k sets of almost the same size. Each one of the *k* sets are partitioned in a way that the ratio of samples of each class in the original data is maintained. Training is done with data from *k-1* sets and tested with one set. To reduce variability, cross validation is carried out using different partitions and the validation results are averaged over all the rounds. For each combination of possible values of parameters, cross validation is carried out. Cross validation can be extensively large if the number of parameters are numerous. Cross validation can be of several types as well. In exhaustive cross validation, the entire data is used for cross validation. On the other hand, in techniques which are non-exhaustive, all ways of splitting the original sample is not done. In this experiment we have used 5-fold cross validation technique.

IV. METHODS

The algorithm proposed in this paper is a supervised machine learning approach. We have chosen the problem of phosphorylation of human proteins to apply our algorithm on. Phosphorylation of Serine (S), Threonine (T) and Tyrosine (Y) have been considered in this experiment because only these amino acid residues can be phosphorylated. In some cases the residue Histidine (H) is phosphorylated but due to the unavailability of sufficient data we could not perform any experiment. For each of the residues, positive data and negative data were extracted from all the human proteins listed in UniProt [12]. Then we removed those proteins which had the residue 'X' or 'B' in it because the physicochemical properties of these residues aren't available in *AAINDEX*.

Before extracting negative data, the database was clustered by using CD-HIT [13] and sequences having similarity greater than 30% were removed. We assume that whether a residue gets phosphorylated depends on its immediate neighborhood. A hypothetical window of odd length is placed which is centered on a prospective site of phosphorylation. The residues lying in this hypothetical window is termed as a segment. If the central residue is phosphorylated then, the segment is called a positive segment otherwise it is termed as a negative segment. For constructing the negative segments we consider only those 'S', 'T' or 'Y' residues which are not phosphorylated. Fig. 1 illustrates how the positive and the negative segments were constructed from the protein sequences. In the positive segment, the residue 'S' has been represented in larger font and for the negative segment the residue 'S' has been represented in small font.

Fig 1. Positive segment and Negative segment

Within a window if there exists any repetitions in either the positive or the negative segments then they are removed. Also those segments are removed which appear both in the positive and the negative segments. Among the indices in *AAINDEX* there are some features, like AVBF000101, which do not have valid data present for all the 20 amino acids. Such indices are initially removed. This led to 531 physicochemical properties. These indices are then normalized between 0 and 1. Now for each residue and each window, the data is prepared by considering each of the 531 physicochemical properties of each amino acid present in the segment. Therefore for window 7 there exists in total (531×7) = 3717 features.

Fig 2. describes the procedure of preparation of training sample from a segment. The entire training sample contains 3717 features. The first seven residues correspond to the scaled value of the first physicochemical property in *AAINDEX* for the 7 amino acid residues in the entire segment. For example in Fig. 2 features 1 to 7 correspond to the physicochemical property 'H ANDN920101', features 8 to 14 correspond to the property 'H ARGP820101' and so on.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 13, October 2016

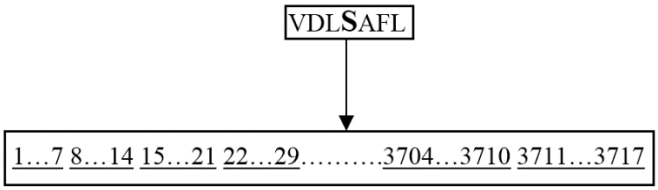

Fig 2. Preparation of data sample from segment

Now we develop a simple scoring technique for each one of the features. A feature will be able to classify data with greater accuracy if the intra class standard deviation is minimized and the inter class mean, between all pairs of classes, be maximized. In our problem, therefore, the aim was to figure out those features for which the sum of standard deviation within each class was very small and sum of the difference between the mean of each pair of classes is very large. We create the following score for each one of the 3717 features.

$$score_i = \frac{\sum_{p=1}^{n}\sum_{q=p+1}^{n}|\mu_{ip}-\mu_{iq}|}{\sum_{j=1}^{n}\sigma_{ij}} \qquad (1)$$

Where μ_{ip} is the mean of the i^{th} feature and p^{th} class. σ_{ij} is the standard deviation of the i^{th} feature and j^{th} class. The 3717 are not the actual physicochemical properties of the amino acids. So we require another scheme to assign scores to each one of the 531 features. For a feature to have a high score the sum of the scores of each residue in the segment should be maximized and standard deviation be minimized. Therefore, each feature was assigned a final score using the formula:

$$final\_score_i = \frac{\sum_{j=1}^{w} score_{ij}}{stddev(score_{ij}, j=1 \text{ to } w)} \qquad (2)$$

Each one of the 531 features now had a *final_score*. The physicochemical properties are now ranked according to their increasing *final_score*. We have experimented with each odd window between 7 and 23. Ideally, the same ranking should have been recorded for each window. But due to noise in data it is impractical to expect 100% similarity in ranking. Hence to select *k* features, we perform an intersection of the order of features from all the windows and choose the *k* topmost features.

V. EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate the performance of our feature selection algorithm, we have performed 5 fold cross validation experiments for each residue and for the window 7. Performance was judged on the basis of area under Receiver Operating Characteristic (ROC) curve metric.

In a Receiver Operating Characteristic curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. ROC curves do not depend on class probabilities, facilitating their interpretation and comparison across different data sets. Originally invented for the detection of radar signals, they were soon applied to psychology and medical fields such as radiology. They are now commonly used in medical decision making, bioinformatics, data mining and machine learning, evaluating biomarker performances or comparing scoring methods.

In the ROC context, the area under the curve (AUC) measures the performance of a classifier and is frequently applied for method comparison. A higher AUC means a better classification. A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 13, October 2016

In AMS3.0 [7], Basu et. al has picked up 10 features from *AAINDEX* to perform prediction of post translation modification in proteins. We have used those features and have performed a 5 fold cross validation with data extracted from human proteins. Alongside we have used our algorithm to select 10 features and have performed the same cross validation experiment in order to compare between the two algorithms.

TABLE I. COMPARISON BETWEEN FEATURES USED IN AMS3.0 AND GSFS

Residue	AMS3.0	GSFS
S	0.845	0.85
T	0.863	0.854
Y	0.796	0.820

Therefore, from Table I it is evident that for residues 'S' and 'Y' GSFS achieves better result than ASM3.0. Saha et.al has adopted unsupervised feature selection in [1]. They have come up with three groups of features each containing 8, 24 and 40 physicochemical features respectively. In order to compare our GSFS algorithm with [1], we also create three groups each with 8, 24 and 40 features respectively. We call these group of features GSFS8, GSFS24 and GSFS40 respectively. The results obtained are summarized in Table II.

TABLE II. COMPARISON BETWEEN HQI AND GSFS

No. of features	S		T		Y	
	HQI	**GSFS**	**HQI**	**GSFS**	**HQI**	**GSFS**
8	0.848	0.843	0.867	0.847	0.797	0.826
24	0.8797	0.8717	0.888	0.884	0.838	0.839
40	0.8759	0.8818	0.886	0.886	0.832	0.845

From the above table it is clear that GSFS performs as well as HQI and in many cases better than HQI itself. It is interesting to note that by increasing of the number of features HQI always does not give better results. For example in case of residue 'Y' the auc_roc score of HQI40 is less than HQI24. This is a subtle indication of inclusion of spurious features. But in case of GSFS always we find a steady increase of auc_roc score with increase of number of features. This indicates that the ranking of the features have been done more efficiently by GSFS algorithm. Below the graph illustrates the performance of the two algorithms for different number of features.

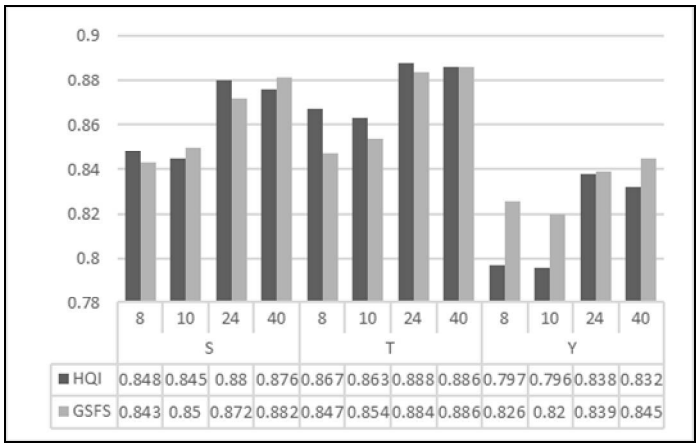

Fig 3. Graph illustrating the comparison of GSFS and HQI with the features used in AMS3.0

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 13, October 2016

VI. CONCLUSION

From the cross validation experiments it is clear that the GSFS algorithm performs as well as the fuzzy clustering in [1]. Designing supervised algorithms is essential because the different sets of features give good results for different classification tasks. In this paper we have proposed a greedy algorithm to perform a ranking of the available features. Then the algorithm tries to combine those features which reappear on the rankings of different windows. The algorithms can be easily extended to work for multi class problems as well.

REFERENCES

[1] I. Saha, U. Maulik, S. Bandyopadhyay, and D. Plewczynski, "Fuzzy clustering of physicochemical and biochemical properties of amino acids," *Amino Acids*, vol. 43, no. 2, pp. 583–594, 2012.
[2] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowl. Inf. Syst.*, vol. 12, no. 1, pp. 95–116, 2007.
[3] K. Jong, E. Marchiori, M. Sebag, and A. Van Der Vaart, "Feature selection in proteomic pattern data with support vector machines," in *Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB'04. Proceedings of the 2004 IEEE Symposium on*, 2004, pp. 41–48.
[4] X. Liu and J. Tang, "Mass Classification in Mammograms Using Selected Geometry and Texture Features, and a New SVM-Based Feature Selection Method."
[5] K. Javed, H. A. Babri, and M. Saeed, "Feature selection based on class-dependent densities for high-dimensional binary data," *Knowl. Data Eng. IEEE Trans.*, vol. 24, no. 3, pp. 465–477, 2012.
[6] D. Plewczynski, S. Basu, and I. Saha, "AMS 4.0: consensus prediction of post-translational modifications in protein sequences," *Amino Acids*, vol. 43, no. 2, pp. 573–582, 2012.
[7] S. Basu and D. Plewczynski, "AMS 3.0: prediction of post-translational modifications," *BMC Bioinformatics*, vol. 11, no. 1, p. 210, 2010.
[8] P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, and D. Plewczynski, "PSP_MCSVM: brainstorming consensus prediction of protein secondary structures using two-stage multiclass support vector machines," *J. Mol. Model.*, vol. 17, no. 9, pp. 2191–2201, 2011.
[9] P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, and D. Plewczynski, "PPI_SVM: Prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables," *Cell. Mol. Biol. Lett.*, vol. 16, no. 2, pp. 264–278, 2011.
[10] B. K. Sriwastava, S. Basu, U. Maulik, and D. Plewczynski, "PPIcons: identification of protein-protein interaction sites in selected organisms," *J. Mol. Model.*, vol. 19, no. 9, pp. 4059–4070, 2013.
[11] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.
[12] The UniProt Consortium, "Update on activities at the Universal Protein Resource (UniProt) in 2013.," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D43–7, Jan. 2013.
[13] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT Suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.