How to Support Prediction of Amyloidogenic Regions - The Use of a GA-based Wrapper Feature Selections

Olgierd Unold
Institute of Computer Engineering, Control and Robotics
Wroclaw University of Technology
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
olgierd.unold@pwr.wroc.pl

***Abstract*—In this paper, we address the problem of predicting the location of amyloidogenic regions in proteins. To support this process we used a genetic algorithm-based wrapper feature subset selection. The wrapper feature subset selection approach is about choosing a minimal subset of features that satisfies an evaluation criterion. We find that most of the machine learning algorithms taken from the WEKA software achieved no worse Accuracy over reduced dataset than over the non-reduced dataset. Moreover, research has confirmed the observations of other researchers, that amino-acids have highly position-dependent propensities.**

***Keywords-Amyloid Proteins; Data Mining; Feature Subset Selection*.**

I. INTRODUCTION

In this paper, we are interested in predicting amyloid proteins. A protein becames amyloid due to an alteration in its secondary structure. A key role in conversion of proteins from their soluble state into fibrillar, beta-structured aggregates, play short sequences, named *hotspots*. Amyloid proteins cause a group of diseases called amyloidosis, such as Alzheimer's, Huntington's disease, and type II diabetes. Symptoms of amyloidosis depend on the organs and tissues amyloid affects.

A laboratory test of a large number of peptides for determining the presence of amyloid protein is in fact theoretically possible, but practically it is not feasible. Therefore, computational methods are commonly used to overcome this limitation. Over the last few years, various computational methods - among existing ones - have been developed to detect these *hotspots* in proteins, like AmylPred [1], Pafig [2], FoldAmyloid [3], and Waltz [4] (for available software dedicated to this task see [5]). However, these methods are very often time consuming algorithms (like 3D Profile method). It is useful, therefore, to use less demanding methods, such as machine learning algorithms, moreover joined with a reduction of the size of the analyzed data.

In this work, we carry an elaborate performance study of different machine learning classification algorithms and feature subset selection (FSS) method applied to Amyloidogenic dataset. All of the algorithms and the wrapper were taken from the Weka machine learning software [6]. The methods over datasets (reduced and non-reduced) were compared in terms of Accuracy.

The remainder of this paper is organized as follows. Section 2 includes state-of-the-art of the problem of predicting amyloidogenic regions, and Section 3 describes Amyloidogenic dataset, FSS method mining the data, and a set of classifiers. Section 4 shows the results obtained, and finally the conclusions are drawn in Section 5.

II. STATE OF THE ART

As established recently [7], there is the strong association between protein fibrils and amyloid diseases, such as Alzheimer's disease, Parkinson's disease, transmissible spongiform encephalopathies, and type II diabetes. It was also observed [8], that amyloids can be formed from short peptide fragments, called hotspots. These strings when exposed to the environment can cause the changeover of native proteins into amyloid state.

Since it is not possible to experimental test all possible protein sequences, several computational tools for predicting amyloid chains have emerged. Most of them are based on physicochemical grounds or structural denominators, like AmylPred [1], Pafig [2], FoldAmyloid [3], and Waltz [4]. However, to our knowledge, no one has used a genetic algorithm-based wrapper feature subset selection method to solve problem under study.

In this paper, we propose a feature subset selection to support predicting amyloid peptides. More over, we are not interested in a time-consuming investigation of physicochemical properties of the amino acids [9], [10], [11], [12] or gaining insight into aggregation propensity [13], [14], [15]. What we are trying to do is to predict amyloidogenic feature of peptide sequence, having no additional knowledge about this sequence. Feature subset selection methods are taken from general-oriented, freely available WEKA software.

III. DATA AND METHODS

A. Data

In our work, we used so-called Waltz amyloidogenic dataset [4]. This is experimentally verified database consisting of 116 amyloidogenic hexapeptides and 162 non-

 ISBN: 978-1-61208-227-1

amyloid-forming hexapeptides. According to its authors, to obtain these data more than 200 peptide sequences were inspected using different structural and biophysical methods. Advantage of Waltz dataset over the others is that it contains experimentally determined structures. Very often amyloid datasets created by various modeling methods – computationally identified – (like the 3D profile methods) [13]) are prone to producing erroneous results. Note that the RosettaDesign potential energy function used in the 3D profile methods is based on heuristic simulated annealing.

B. Classifiers

The experiments were conducted comparing the classification Accuracy of 13 classification methods implemented in the Weka software. Here we briefly list the classifiers that we used:

- *Naive Bayes* and *BayesNet* – classifiers based on the Bayesian Theorem in which it is assumed that the attributes have equal weight and are conditionally independent,
- *Support Vector Machine* – algorithm trying to find a hypersurface in the space of possible inputs,
- *C4.5*, *Random Tree*, *REPTree*, *RandomForest*, *ADTRee* – methods creating a hierarchy of nodes, each associated with a decision rule on one attribute. ADTree creates alternating decision trees, RandomTree builds a tree considers a given number of random features at each node, RandomForest builds random forests using Breiman's algorithm in which multiple random trees vote on an overall classification for the given set of inputs. REPTree uses reduced-error pruning to speed up a learning process, C4.5 algorithm improves Quinlan's method for decision tree induction,
- *JRip* – classifier generating rules, which can transformed from or in decision trees. JRip is the WEKA version of RIPPER, which is a rule-based learner that builds a set of rules that identify the classes while minimizing the amount of error. The error is defined by the number of training examples misclassified by the rules,
- *MultiLayer Perceptron* – kind of simple neural network classifier, in which backpropagation algorithm calculates connection weigths given a fixed network structure,
- *KStar* – an instance-based classifier using an Entropic Distance Measure. It provides a consistent approach to handling of symbolic attributes, real valued attributes and missing values,
- *AdaBoost* – one of the most popular boosting algorithms. Boosting is an iterative method in which new model is efected by the performance of those built previously. This is achieved by assigning proper weights to learning instances in each iteration,
- *END* – a meta-classifier for handling multi-class datasets. The main idea of meta-classification is to represent the judgment of each classifier (SVM-based) for each class as a feature vector, and then to re-classify again in the new feature space. The final decision is made by the meta-classifiers instead of just linearly combining each classifiers judgment.

More information on implementing in the Weka software classifiers is presented in [6].

The quality of our predictions was evaluated using the commonly used standard value Accuracy, which is measured by the number of correct results, the sum of true positives and true negatives, in relation to the number of tests carried out

$$Accuracy = (\frac{TruePositives + TrueNegatives}{Total}) \times 100 \quad (1)$$

where True Positives are correctly (i.e., as amyloidegenic peptides) recognized positive examples, True Negatives - correctly recognized negatives (i.e., as non-amyloidogenic ones).

C. GA-based Wrapper Feature Selection

Feature selection methods can be put into two main categories from the point of view of a methody output. One category, called filter approach, comprises methods ranking features according to the same evaluation criterion; the other, called the wrapper approach, consists of methods choosing a minimum subset of features that satisfies an evaluation criterion.

It was proved that the wrapper approach produces the best results out of the feature selection methods [16], although this is a time-consuming method since each feature subset considered must be evaluated with the classifier algorithm. In the wrapper method, the attribute subset selection algorithm exists as a wrapper around the data mining algorithm and outcome evaluation. The induction algorithm is used as a black box. The feature selection algorithm conducts a search for a proper subset using the induction algorithm itself as a part of the evaluation function. GA-based wrapper methods involve a genetic algorithm (GA) as a search method of subset features.

GA is a random search method, effectively exploring large search spaces [17]. The basic idea of GA is to evolve a population of individuals (chromosemes), where individual is a possible solution to a given problem. In case of searching the appropriate subset of features, a population consists of different subsets evolved by a mutation, a crossover, and selection operations. After reaching maximum generations, algorithms returns the chromosome with the highest fitness, i.e. the subset of attributes with the highest Accuracy.

 ISBN: 978-1-61208-227-1

IV. EXPERIMENTAL RESULTS

The comparison was performed using a recently published Amyloidogenic dataset, composed by 116 hexapeptides known to induce amyloidosis and by 162 hexapeptides that do not induce amyloidosis [4]. In our experiments, we randomly split the Amyloidogenic database into 10 equally folds, and use a 10-fold cross validation method to determine the classification Accuracy.

A k-fold cross validation (k-fold cv) is a well-established statistical method of evaluating a learner, combining training and validation phases [18]. In k-fold cv the data is partioned into k folds, and next subsequently k iterations of learning and testing are performed such that within each iteration a different fold of the data is held-out for validation while the remaining $k-1$ folds are used for learning.

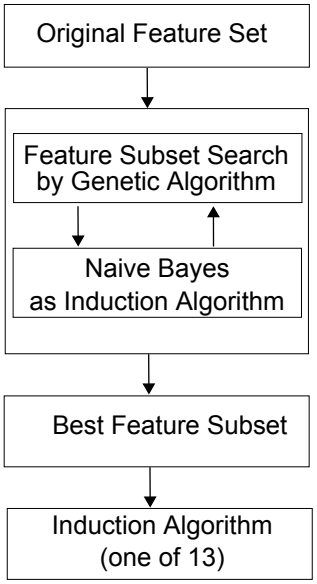

Figure 1. GA-based wrapper feature selection with Naive Bayes as an induction algorithm evaluating feature subset

We employ 13 commonly used machine learning algorithms: BayesNet, NaiveBayes, MultiLayerPerceptron (MLP), Support Vector Machine (SMO), KStar, AdaBoost, END, JRip, C4.5, Random Tree, REPTree, RandomForest, ADTree and GA-based wrapper approach for feature selection. GA-based wrapper methods involve a genetic algorithm as a search strategy of subset features, and one of the machine learning method as an induction algorithm, in this paper NaiveBayes (see Fig. 1). The study [19] noted that no significant difference exists between results achieved by various induction algorithms used in a GA-based wrapper method. All of the classification algorithms and the wrapper were taken from the Weka software [6], all of them used default parameters.

Table 1 summarizes the performances of the 13 compared methods over reduced (denoted as a Dataset 1-3-5), and non-reduced Amyloidogenic dataset (Dataset 1-2-3-4-5-6). Ten of the thirteen methods gained better results over reduced dataset, although the results were not confirmed statistically. What is interesting, the feature selection method chooses only three from six amino acids as important in hexapeptide, in positions 1, 3, and 5. Note that such observations were also made in laboratory experiments. Maurer-Stroh et al. [4] recorded the strong position-specific tendencies of the different amino acids for forming amyloid structures.

Table I
THE PERFORMANCES IN TERMS OF ACCURACY OF THE MACHINE LEARNING METHODS OVER REDUCED AND NON-REDUCED AMYLOIDOGENIC DATASET. THE HIGHER ACCURACY IN A ROW IS INDICATED IN BOLD.

Method	Dataset 1-3-5	Dataset 1-2-3-4-5-6
BayesNet	**68.02**	65.57
NaiveBayes	**66.93**	65.57
MLP	64.76	**72.34**
SMO	68.37	**73.81**
KStar	63.69	**65.50**
AdaBoost	**69.80**	68.37
END	**64.03**	60.81
JRip	**69.06**	68.74
ADTree	**73.77**	69.81
C4.5	**64.03**	60.81
RandomTree	**66.55**	65.91
REPTree	**66.90**	66.28
RandomForest	**66.56**	65.15

V. CONCLUSION

The problem of predicting the amyloidogenic regions in proteins was addressed. Our analysis showed that the use of feature subset selection can support efficiently this task. In most cases machine learning methods have achieved better results over reduced dataset. In addition, methods processed twice smaller learning set.

It is worth noticing that the overall best results have been gained by Support Vector Machine over non-reduced data (73.81 % of Accuracy), and Alternating Tree but over reduced dataset (73.77 %). If SVM is quite often used in prediction different regions in protein chains [2], the ADTree gives interpretable and understandable by human results.

The performed computational experiments confirm also laboratory studies over proteins, in which the strong position dependency of residues are observed.

REFERENCES

[1] K. Frousios, V. Iconomidou, C. Karletidi, and S. Hamodrakas, "Amyloidogenic deteminants are usually not buried," *BMC Structural Biology*, vol. 9, p. 44, 2009.

[2] J.Tian, N. Wu, J. Guo, and Y. Fan, "Prediction of amyloid fibril-forming segments based on a support vector machine," *BMC Bioinformatics*, vol. 10 (Suppl 1):S45, 2009.

 ISBN: 978-1-61208-227-1

[3] S. Garbuzynskiy, M. Lobanov, and O. Galzitskaya, "An introduction to variable and feature selection," *Bioinformatics*, vol. 26, pp. 326–332, 2010.

[4] S. Maurer-Stroh, M. Debulpaep, N. Kuemmerer, M. L. de la Paz, I. Martins, J. Reumers, K. Morris, A. Copland, L. Serpell, L. Serrano, J. Schymkowitz, and F. Rousseau, "Exploring the sequence determinants of amyloid structure using position-specific scoring matrices," *Nat Methods*, vol. 7, pp. 237–242, 2010.

[5] S. Hamodrakas, "Protein aggregation and amyloid fibril formation prediction software from primary sequence: Towards controlling the formation of bacterial inclusion bodies," *FEBS Journal*, vol. 278, no. 14, pp. 2428–2435, 2011.

[6] I. Witten, E. Frank, and M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques. Third edition*. Morgan Kaufmann, 2011.

[7] L. Goldschmidt, P. Teng, R. Riek, and D. Eisenberg, "Identifying the amylome, proteins capable of forming amyloid-like fibrils," *PNAS*, vol. 107(8), pp. 3487–3492, 2010.

[8] O. Galzitskaya, S. Garbuzynskiy, and M. Lobanov, "Prediction of amyloidigenic and disordered regions in protein chains," *PLoS Computational Biology*, vol. 2(12), p. e177, 2006.

[9] G. Tartaglia, A. Cavalli, R. Pellarin, and A. Caflisch, "Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences," *Protein Sci*, vol. 14(10), pp. 2723–2734, 2005.

[10] K. DuBay, A. Pawar, F. Chiti, J. Zurdo, C. Dobson, and M. Vendruscolo, "Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains," *J Mol Biol*, vol. 341(5), pp. 1317–1326, 2004.

[11] A. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, and L. Serrano, "Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins," *Nat Biotechnol*, vol. 22(10), pp. 1302–1306, 2004.

[12] S. Idicula-Thomas and P. Balaji, "Understanding the relationship between the primary structure of proteins and their amyloidogenic propensity: clues from inclusion body formation," *Protein Eng Des Sel*, vol. 18(4), pp. 175–180, 2005.

[13] M. Thompson, S. Sievers, J. Karanicolas, M. Ivanova, D. Baker, and D. Eisenberg, "The 3D profile methods for identifying fibril-forming segments of proteins," *Proc Natl Acad Sci USA*, vol. 103, pp. 4074–4078, 2006.

[14] S. Yoon and W. Welsh, "Detecting hidden sequence propensity for amyloid fibril formation," *Protein Sci*, vol. 13(8), pp. 2149–2160, 2004.

[15] M. L. D. L. Paz, K. Goldie, J. Zurdo, E. Lacroix, C. Dobson, A. Hoenger, and L. Serrano, "De novo designed peptide-based amyloid fibrils," *Proc Natl Acad Sci USA*, vol. 99(25), pp. 16 052–16 057, 2002.

[16] X. Zhiwei and W. Xinghua, "Research for information extraction based on wrapper model algorithm," in *2010 Second International Conference on Computer Research and Development*, Kuala Lumpur, Malaysia, 2010, pp. 652–655.

[17] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA, 1989.

[18] F. Mosteller and J. Turkey, *Data Analysis, Including Statistics*, ser. Handbook of Social Psychology. Addison-Wesley, Reading, MA, 1968.

[19] O. Unold, M. Dobrowolski, H. Maciejewski, P. Skrobanek, and E. Walkowicz, "A GA-based wrapper feature selection for animal breeding data mining," *Lecture Notes in Computer Science*, vol. 7209, pp. 200–209, 2012.

 ISBN: 978-1-61208-227-1