RESEARCH ARTICLE OPEN ACCESS

Protein Sequence Classification Using Feature Selection

K. Radha MCA[1], C. Akila MCA., MPhil[2]
Research Scholar, Sree Saraswathi Thyagaraja College,
Affiliated to Bharathiar University, Coimbatore
Head, PG Department of Computer Science,
Sree Saraswathi Thyagaraja College,
Pollachi-TamilNadu
India

ABSTRACT

The sheer volume of data today and its expected growth over the next years are some of the key challenges in data mining and knowledge discovery applications. Besides the huge number of data samples that are collected and processed, the high dimensional nature of data arising in many applications causes the need to develop effective and efficient techniques that are able to deal with this massive amount of data. In addition to the significant increase in the demand of computational resources, those large datasets might also influence the quality of several data mining applications (especially if the number of features is very high compared to the number of samples). As the dimensionality of data increases, many types of data analysis and classification problems become significantly harder. This can lead to problems for both supervised and unsupervised learning. Dimensionality reduction and feature (subset) selection methods are two types of techniques for reducing the attribute space. While in feature selection a subset of the original attributes is extracted, dimensionality reduction in general produces linear combinations of the original attribute set.

Keywords:- Feature Selection, Feature Reduction, Feature Prediction.

I. INTRODUCTION

The tremendous improvements in techniques for collecting, storing and transferring large volumes of data have also increased the volume of data for knowledge discovery and data mining applications. Data grow not only due to the number of data samples available, but also due to the increasing number of candidate features for various application areas. Not only the effort and computational cost of data mining applications grow with increasing dimension of data. Bioinformatics has been an active area of research for the last three decades and is continuously gaining thoughtful attention from computer scientists and biologists research community. The objectives of bioinformatics were to store and manage the biological data and develop sophisticated computational tools that are helpful in the analysis and modeling.

The data generally consists of deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins. The most fundamental element of any living organism is proteins. It comprises amino acids that carry out an important role in cell functions including nutrient transportation, metabolism regulation, and muscle building. A protein may adapt four different types of conformations due to some structural changes in order to perform functions inside the cell in the human body [3]. Every unknown protein needs annotation to know its structure and function, while the speed of the in vitro experiments is lessened quite a bit as more and more novel sequences are added constantly in the protein databases. However, the experimental methods are finding difficulties in annotating new proteins as they are very labor intensive and take a long time. The homology-based approaches also have been utilized to predict the function of unannotated proteins by finding the sequence homology found between sequences in the databases. Two main categories of sequence homology based approaches are alignment-based and alignment-free. Alignment-based models depend on single or multiple alignments to construct different types of models. Recently, techniques like basic local alignment search tool (BLAST), FASTALL (FASTA), and HIDDEN MARKOV MODELS (HMM) were the most reliably used alignment-based traditional methods for the analysis of both protein and DNA sequences.

The results of protein BLAST show which segment or part of the protein sequence has more matches with the already available protein sequences in the database. BLAST uses the heuristic algorithm to measure the statistical significance of matched sequences in order to find similarity among them, while FASTA exploits local sequence alignment to find similar sequence using heuristic search in the database. HMM is a probabilistic model or simple Bayesian model with hidden states. An HMM model is constructed for each family separately. The results of the aligned sequences of amino acid residues are generally represented as rows of a matrix. Generally, obtaining an efficient multiple alignments looks impossible when the sequences do not have enough similarity between them. Sequence alignment programs use a scoring matrix such as point accepted mutation (PAM) and BLOCKS SUBSTITUTION MATRIX (BLOSUM) to generate a score for the alignment. Some limitations of alignment-based approaches are as follows.

(i) Alignment-based techniques undergo performance degradation on sequences having very weak or low similarity among them.

(ii) Alignment-based techniques are heuristic in nature and thus are computationally expensive and take a long time on large datasets.

(iii) Alignment-based techniques assume that contiguity is preserved within homologous segments, but this may not be accurate in genetic recombination.

The limitations of the alignment-based protein classification have been removed by the alignment-free classification techniques. These techniques obtain different descriptors from each protein sequence (like the composition of amino acid, amino acid frequencies, and different chemical properties).

II. MOTIVATION

As the dimensionality of the feature space increases, many types of data analysis and classification also become significantly harder, and, additionally, the data becomes increasingly sparse in the space it occupies which can lead to big difficulties for both supervised and unsupervised learning. This phenomenon (known as the curse of dimensionality) is based on the fact that high dimensional data is often difficult to work with. A large number of features can increase the noise of the data and thus the error of a learning algorithm, especially if there are only few observations (i. e., data samples) compared to the number of features. Feature selection and dimensionality reduction methods (summarized as feature reduction methods) are two techniques that aim at solving these problems by reducing the number of features and thus the dimensionality of the data.

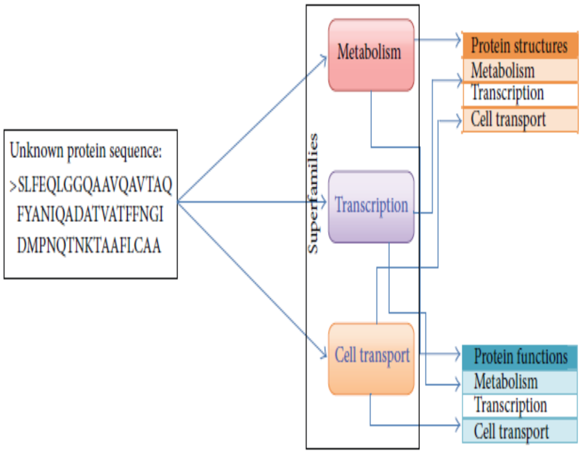

Figure-1: Prediction of the structure or function of an unknown protein.

Figure-1 explains the concept of the determination of the structure and function of any protein exclusively from the primary amino acid sequence. Moreover, Figure 1 demonstrates that, for a given unknown sequence, the classification technique investigates with which super family the new protein sequence belongs based on similarity with the existing sequences. In the figure, only three yeast sample super families, namely, metabolism, transcription, and cell transport, were shown. The unknown sequence may belong to one of the three super families based on the structure and function similarity. The high dimensionality of biological data creates several crucial problems for the researchers during the implementation of machine learning based approaches during the analysis and modeling of extremely large amounts of sequence data. Many feature selection techniques have been introduced but still there is a need for a technique that can select statistically significant features for each protein sequence. The feature reduction would increase classification accuracy by removing the redundant or unnecessary features and also decrease the running time of classification algorithms. The automatic classification mechanism saves long time required for the experiments and the expenses of costly biological tests in laboratories.

III. RELATED WORK

Jeong et al. introduced a feature extraction method based on the position specific scoring matrix (PSSM) to extract features from a protein sequence. The PSSM consisted of four components: position, probe, profile, and consensus. The authors defined four feature sets from the PSSMs. Feature set number 1 was obtained by dividing a protein sequence of any length into 20 equal sized blocks. Feature set number 2 considered domains having the same conservation ratio. Feature set number 3 extracted the physicochemical properties of the probing residues obtained from feature set number 2. Finally, feature set number 4 was proposed which consisted of all three feature sets. The total number of features investigated by this technique comprised a combination of four feature sets. Afterwards, the authors used four classifiers for the evaluation of classification technique: the naive Bayesian (NB), support vector machine (SVM), decision tree (DT), and random forest (RF). Three yeast super families (i.e., metabolism, transcription, and cellular transport) sequences were used as a training and test dataset. The maximum classification accuracy obtained was 72.5%. The accuracy was low due to a high misclassification rate.

However, the accuracy could be further improved by extracting more relevant features from the protein sequence.

Mansoori et al. extracted features from a protein sequence using 2 grams and a 2-gram exchange group from the training and test data. The distance-based feature ranking method was used for the selection of the best and most appropriate features. A SGERD-based classifier (steady state genetic algorithm for extracting fuzzy rules from data) was used to create fuzzy rules. Five super families were considered in the experiments: globin, insulin, kinase, ras, and trypsin. These rules were then used for the classification of the protein sequences into super families. The authors proposed a method that reduced the classification time from 79 to 51 minutes, while the classification accuracy was 96.45%. The time required for the classification could be further reduced and there would also be fewer chances that similar 2 grams would occur in unrelated sequences. Further improvement could also be made in the classification accuracy and the running time of the classification algorithm by the application of an appropriate feature selection technique.

Bandyopadhyay proposed a method that used a 1-gram technique for feature encoding. The feature size was comprised of 20 amino acids. The extracted feature reflected the probability with which each amino acid occurred in any protein sequence. The authors proposed a variable length fuzzy genetic clustering algorithm to find prototypes for each super family. For classification of protein sequences to relevant super families, the nearest neighbor algorithm was employed. Three super families, globin, ras, and trypsin, were utilized in the experiments. The classification accuracy obtained on the mentioned dataset was 81.3%.

The classification accuracy can be enhanced using highly informative and more relevant features to describe a variable-length protein sequence. In addition to the above works, in [26], the authors used different physicochemical properties to represent the features of a protein sequence. Only the distinguished and invariant features were used in the experiments. In the experiments, three super families, such as esterase, lipase, and cytochrome, was investigated. The extracted features were given as input to the feed-forward, probabilistic neural network and radial basis function neural network. The probabilistic neural network showed accuracy of 90.6% on three super families: esterase, lipase, and cytochrome. The classification accuracy might be increased by introducing a feature selection technique that has good discrimination power during classification. Leslie et al. proposed a spectrum kernel to measure the sequence similarity between protein sequences. The technique considered subsequences of k length amino acids (k-spectrum kernel) as a feature vector. The feature vector space obtained from the spectrum kernel was then passed to a support vector machine for classification of protein sequences into their relevant classes.

IV. FEATURE SELECTION

Feature selection techniques do not alter the original representations of features, but select a subset of them. Hence, these techniques preserve the original semantics of the features, offering the advantage of interpretability by a domain expert. In theory, the goal is to find the optimal feature subset that maximizes the scoring function above. The selection of features (or subsets of features) should be performed on the training set only, the test set is then used to validate the quality of the selected features (subsets). Feature subset selection approaches are categorized into three main groups: filter methods, wrapper methods and embedded approaches. Filter methods rely on general characteristics of the training data to evaluate and select subsets of features without involving a learning algorithm. Contrary to that, wrapper approaches use a classification algorithm as a black box to assess the prediction accuracy of various subsets. The last group, embedded approaches, performs the feature selection process as an integral part of the machine learning algorithm.

Filter methods are classifier agnostic, no-feedback, pre-selection methods that are independent of the machine learning algorithm to be applied. Filter methods can further be divided into Univariate and multivariate techniques. Univariate filter models consider one feature at a time, while multivariate methods consider subsets of features together, aiming at incorporating feature dependencies. Univariate filter methods are referred to as single variable classifiers, and multivariate filter methods are grouped together with wrapper methods and embedded methods and referred to as variable subset selection methods. Univariate filter method consider features separately and usually make use of some scoring function to assign weights to features individually and rank them based on their relevance to the target concept. This procedure is commonly known as feature ranking or feature weighting. A feature will be selected if its weight or relevance is greater than some threshold value.

V. DIMENSIONALITY REDUCTION

Dimensionality reduction (DR) refers to algorithms and techniques which create new attributes as (often linear) combinations of the original attributes in order to reduce the dimensionality of a dataset. Rather than selecting a subset of the features, these techniques involve some type of feature transformation and aim at reducing the dimension such that the representation is as faithful as possible to the original dataset, but with a lower dimension and removed redundancy. Because the new attributes are combinations of the original ones, the transformation process is also referred to as feature construction or feature transformation. This process of constructing new features can be followed by or combined with a feature subset selection process {the original feature set is first extended by the newly constructed features and then a subset of features is selected.

The study shows that adding newly computed features to the original attributes can increase the classification results achieved with these feature sets more than replacing the original attributes with the newly computed features. In contrast to many feature selection methods, dimensionality reduction techniques usually map the data to lower dimension in an unsupervised manner, i. e., the class labels are not considered, just the explanatory variables are used.

VI. RESULTS

Table-1 shows, on the **non-plant** data set, the performance of feature selection on fixed length as well as variable length *k*-gram representations, where the data size is set to 2^{22}. As seen in the table, the performance of feature selection trained on fixed length *k*-gram representations is worse than that of feature selection trained using variable length *k*-gram representations, with *k* ranging from 1 to 4 resulting in the highest performance (the representation is denoted by (1-4)-grams). The performance of feature selection trained on fixed-length *k*-gram representations is expected to be worse than that of their counterparts trained on variable length *k*-gram representations, as protein sequence motifs have usually variable length. The performance of feature selection trained using variable length *k*-gram representations increases as we add more dependencies in the data (i.e., larger values of *k*), but starts decreasing as *k* becomes greater than 4, which may be due to over fitting.

Bag of fixed or variable length k-grams	**Accuracy (%)**	**No. of Features**
1- Gram	71.20	20
2- Grams	70.84	401
3- Grams	78.81	7990
4- Grams	79.01	145658
(1-2) – Grams	70.57	423
(1-3) – Grams	79.55	8417
(1-4) – Grams	82.83	155017

Table-1: Comparison of fixed-length with variable-length *k*-gram representations.

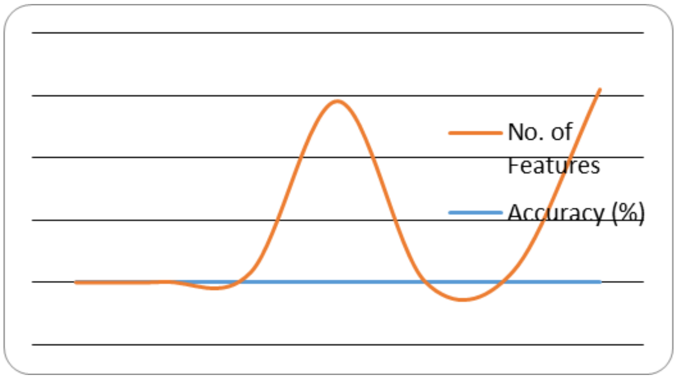

Figure-2 : Comparison of fixed-length with variable-length *k*-gram representations.

The number of variable length *k*-grams, for *k* ranging from 1 to 4, is 155,017. Feature Selection eliminates the need for storing the vocabularies in memory by implicitly encoding the mapping from strings to integers into a hash function. We conclude that feature hashing is very effective on prohibitively high-dimensional *k*-gram representations, which would otherwise be impractical to use. Because (1-4)-gram representation results in the highest performance, we used it for subsequent experiments.

VII. CONCLUSION

The proposed feature subset selection technique uses a threshold to select the highly informative and important features. The results of the technique were validated through the well-recognized classification/learning algorithms. The protein sequences of three different datasets have been effectively classified into relevant super families with substantially high classification accuracy. The introduced classification method is alignment-free, simple, fast, and reliable. This technique of feature selection and classification would be useful in machine learning and bioinformatics in reducing the high dimensionality of data during the prediction of the structure or function of unknown protein sequences.

REFERENCES

[1] N. M. Luscombe, D. Greenbaum, and M. Gerstein, "What is bioinformatics? A proposed definition and overview of the field," Methods of Information in Medicine, vol. 40, no. 4, pp. 346–358, 2001.

[2] D. R. Bentley, "The human genome project—an overview," Medicinal Research Reviews, vol. 20, pp. 189–196, 2000.

[3] J.-M. Claverie and C. Notredame, Bioinformatics for Dummies, 2nd edition, 2007.

[4] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," Proceedings of the National Academy of Sciences of the United States of America, vol. 85, no. 8, pp. 2444– 2448, 1988.

[5] W. Pearson, "Finding protein and nucleotide similarities with FASTA," Current Protocols in Bioinformatics, chapter 3, unit3.9, 2004.

[6] S. F. Altschul, T. L. Madden, A. A. Schaffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Research, vol. 25, no. 17, pp. 3389–3402, 1997.

[7] S. F. Altschul, W.Gish,W. Miller, E.W. Myers, and D. J. Lipman, "Basic local alignment search tool," Journal of Molecular Biology, vol. 215, no. 3, pp. 403–410, 1990.

[8] W. R. Pearson, "Using the FASTA program to search protein and DNA sequence databases," Methods in Molecular Biology, vol.25, pp. 365–389, 1994.

[9] T. Plotz and G. A. Fink, "A new approach for HMM based protein sequence family modeling and its application to remote homology classification," in Proceedings of the IEEE/SP 13th Workshop on Statistical Signal Processing, pp. 1008–1013, Bordeaux, France, July 2005.

[10] K. Karplus, C. Barrett, and R.Hughey, "HiddenMarkovmodels for detecting remote protein homologies," Bioinformatics, vol. 14, no. 10, pp. 846–856, 1998.