35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Support Vector-Based Takagi-Sugeno Fuzzy System for the Prediction of Binding Affinity of Peptides

Volkan Uslan and Huseyin Seker*

Abstract— **High dimensional, complex and non-linear nature of the post-genome data often adversely affects the performance of predictive models. There are two methods that have been widely used to model such non-linear systems, namely Fuzzy System (FS) and Support Vector Machine (SVM). FS is good at modelling uncertainty and yielding a set of interpretable IF-THEN rules, but suffers from the curse of dimensionality whereas SVM is a method that has been shown to effectively deal with large number of dimensions leading to better generalization ability. In this paper, a hybrid system is therefore proposed to improve FS with the aid of SVM-based regression method and successfully applied to the prediction of binding affinity of peptides, which is regarded as one of the most complex modelling problems in the post-genome era due to the diversity of peptides discovered. The proposed hybrid method yields comparatively better results than what has been presented in the recently published papers, therefore can also be considered for other bioinformatics applications.**

I. INTRODUCTION

A peptide is a sequence with a small number of amino acids that are linked together by a peptide bond. Peptide binding plays vital roles in regulating cell signaling and understanding the mechanisms of protein-peptide interactions. As there are many thousands of peptides, identification of binding and its affinity between proteins and peptides requires laborious biological process and is time-consuming. Therefore, there is a need to develop a computational predictive model that is capable of determining the binding and its affinity. As the peptides can be represented with a few thousands of descriptors, this curse of dimensionality makes this prediction process a complex task. In addition, it gets much harder due to availability of very small number of peptides, affinity of which have been experimentally obtained [1].

Computational methods are utilised for building prediction models in many biological problems. One of the methods utilised to model non-linear systems is the fuzzy systems [2] that have been shown to be more capable of modeling uncertain and imprecise knowledge and forming a structure that can represent human reasoning in various applications. Although there are different fuzzy systems, Takagi-Sugeno (TS) is commonly utilised for modeling complex systems [3]. Although there are many methods proposed to model TS fuzzy system (TSFS), general approach is to keep the premise parameters constant whereas values of the consequent parameters are computed by the least square estimation. There are methods that have been explored for addressing to the problems in the least square estimation. One of the methods is support vector regression (SVR) [4] that has been shown to be an efficient and robust method and provides high generalizability and performance. Applications of SVR have demonstrated considerably better modeling of various non-linear systems. Therefore in this paper this concept is incorporated with TSFS to better train the consequent part of the TSFS. Recently, general frameworks to integrate fuzzy systems with the support-vector based methods were presented [5], [6]. However, their applications were only on a very small number of features as similar to other FS applications but did not take into account of predictive problem with very large number of attributes. For the premise part of FS, fuzzy clustering has been used to approximate the membership functions that characterize each fuzzy set involved in the rule-base [7].

In this paper, a hybrid system, called support-vector based Takagi-Sugeno fuzzy system (TSFS-SVR) is proposed and applied to quantitatively predict peptide binding affinities in order to show its robustness. The rest of the paper starts by reviewing Takagi-Sugeno fuzzy system in section II followed by the determining the structure of the fuzzy system (section III). The characteristics of the peptide binding affinity datasets are detailed in section IV. Performance measurements of the prediction models are presented in section V. Results and discussion are provided in Section VI, and finally the conclusions are presented in section VII.

II. TAKAGI-SUGENO FUZZY SYSTEM

Takagi-Sugeno (TS) fuzzy systems have a structure of fuzzy rules such that the premise is constituted by fuzzy sets and the rule outputs are determined as a linear function of input variables. The TS model rules are defined as conditional statements that are presented by using a linear function in the consequent part. A fuzzy rule-base with n input variables (x_1, x_2, ..., x_n), r rules and a single output variable y can be written as:

$$\begin{aligned} &R_r : \text{IF } x_1 \text{ is } A_{1r} \text{ AND } x_2 \text{ is } A_{2r} \text{ ... AND } x_n \text{ is } A_{nr} \\ &\text{THEN } z_r \;=\; f(x_1, x_2, ..., x_n) \end{aligned} \tag{1}$$

where A_{nr} is a fuzzy set for the variable n and rule r, generally represented by a membership function, and z_r is a linear function in the consequent part and can be defined as:

$$z_r = f(x_1, x_2, ..., x_n) = m_0 + \sum_{i=1}^{n}(m_i x_i) \tag{2}$$

[1] The authors are with the Bio-Health Informatics Research Group, Centre for Computational Intelligence, Faculty of Technology, De Montfort University, Leicester, LE1 9BH, UK, email: vuslan@dmu.ac.uk, hseker@dmu.ac.uk

* Corresponding author (hseker@dmu.ac.uk)

978-1-4577-0216-7/13/$26.00 ©2013 Crown

where m_0, m_1, m_2, ..., m_n are the coefficients of input parameters (x_1, x_2, ..., x_n).

In the TS model each rule generates a crisp output and then the final output is obtained by aggregating all the rule outputs. This process is called defuzzification, and the weighted average defuzzification value y can be defined as:

$$y = \sum_{i=1}^{r} \overline{\nu_i} z_i \tag{3}$$

$$\overline{\nu_i} = \frac{\nu_i}{\sum\limits_{k=1}^{r} \nu_k} \tag{4}$$

where ν_i and $\overline{\nu_i}$ are the firing strength and normalized firing strength of the fuzzy rule, respectively and ν_i is determined by using a t-norm operator that can be defined as:

$$\nu_i = \prod_{j=1}^{n} \mu(x_j) \tag{5}$$

where $\mu(x_j)$ is the membership degree of input variable x_j.

The fuzzy sets (e.g., A_{ij}) can be described by any form of membership functions, in which case Gaussian membership function is used and can be defined as:

$$\mu(x_j) = e^{-\frac{(x_j - c_{ij})^2}{2(\sigma_{ij})^2}} \tag{6}$$

where c and σ are centre and standard deviation, respectively.

III. Determining the Structure of The Fuzzy System

A. Determining the Consequent Parameters: Support Vector Regression

Support Vector Machine (SVM) is a statistical learning architecture based on the structural risk minimization [8]. SVM learning algorithm finds the optimal separating hyperplane by training a classifier for a given training data. The optimal separating hyperplane is the one that maximizes the margin between two classes. SVMs can be generalized to regression using its linear model. Other than traditional square error loss function, the ϵ-insensitive loss function is used in SVR. This chosen error function tolerates errors up to ϵ. One other advantage of using this error function is its tolerance against noise. Similar to the soft margin hyperplane in SVMs, slack variables are used for deviations out of the ϵ-region. SVR searches for a linear function $h(x)$:

$$h(x) = w^T x + b. \tag{7}$$

which is constrained to the following mathematical expressions:

$$\min \frac{1}{2}\|w\|^2 + C \sum (\xi_+ + \xi_-). \tag{8}$$

subject to:

$$\begin{array}{l} y' - (w^T x + b) \leq \epsilon + \xi_+ \\ (w^T x + b) - y' \leq \epsilon + \xi_- \\ (\xi_+, \xi_-) \geq 0 \end{array} \tag{9}$$

where two types of slack variables ξ^+, ξ^- are used to optimise the parameters and w and b represent the coefficients of the weight vector of the linear expression. The parameter C is a pre-specified value and works as a regularization factor between minimizing w and up to the value which deviations greater than ϵ can be tolerated. Similar to the classification problem, a certain training instances are chosen to be support vectors. Then, the weighted sum of the support vectors are used to define the regression and adequately model data.

A common method used to compute values of the consequent parameters of TSFS is the least squares estimation [9]. Given the support vector regression concept with a linear kernel, this can be potentially utilized to compute values of the consequent parameters of TSFS. SV regression part was implemented using LIBSVM library [10].

The variables ($\overline{\nu_i}$, $\overline{\nu_i}x_{i1}$, $\overline{\nu_i}x_{i2}$, ..., $\overline{\nu_i}x_{in}$) defined using the normalized firing strength (Eq. 4), form inputs to SVR to derive w parameters that correspond to the consequent parameters in TSFS. Finally, SV-based TSFS can be formulated as:

$$z'_r = f(x_1, x_2, ..., x_n)' = w_{0r} + \sum_{i=1}^{n} (w_{ir} x_i) \tag{10}$$

$$y' = \sum_{i=1}^{r} (\overline{\nu_i} z'_i + \frac{b}{r}) \tag{11}$$

where y' now represents the formulation of the SVR-based TSFS.

B. Determining the Antecedent Parameters: Fuzzy C-Means

Fuzzy c-Means (FCM) algorithm partitions the dataset into a number of clusters by assigning a degree of membership for each data object to all the clusters [11]. The FCM algorithm is constrained to minimize the objective function by measuring the squared sum of distance between data objects and cluster centres in any inner product norm, and the membership of data objects with a weight exponent.

The membership values and cluster prototypes obtained from the fuzzy clustering algorithm can be used to approximate the membership functions. Moreover, the fuzzy sets involved in the rules are fully characterised by their membership functions. Each partition provides information such as centroid of a cluster, standard deviation of data objects, all which can be used to derive membership functions. The clusters and their parameters form the premise part of TSFS and number of clusters is used to determine the number of rules.

C. Reducing the High Dimensionality: Feature Selection

Feature selection is the process of choosing a subset of features used to improve the efficiency of the system or the model. Features are reduced to the least number of dimensions possible that yields higher accuracy and performance. Although there are various types of feature selection methods used in the post-genome data analysis, an unsupervised feature selection method, namely multi-cluster feature selection (MCFS), is used [12] in order to

TABLE I

GENERAL CHARACTERISTICS OF THE DATASETS USED FOR THE PREDICTION OF PEPTIDE BINDING AFFINITY

Datasets	Number of Peptide Sequences		Number of Peptide Sequence Descriptors
	Training	Testing	
Task 1	89	88	5787
Task 2	76	76	5144
Task 3	133	133	5787

TABLE II

PEPTIDE BINDING AFFINITY CHARACTERISTICS

Datasets	Training				Testing			
	Min	Max	Mean	Std	Min	Max	Mean	Std
Task 1	2.94	8.65	5.41	1.01	3.13	8.17	5.41	0.95
Task 2	5.01	8.34	7.55	0.77	5.01	8.40	7.58	0.74
Task 3	4.30	8.77	7.08	0.82	5.08	8.96	7.10	0.80

make sure that the features stratified are independent of any predictive model. This method uses information contained in eigenvectors by solving the generalised eigen-problem to preserve the multi-cluster structure of the data. This approach has also been shown to be able to deal with large number of attributes, which is very common and a key problem in the post-genome data. The reduced feature subset will be used as input variables of the rule-based fuzzy system. This low dimensional structure will help eliminate noise in the dataset and will provide more robust rule-base to model a non-linear system.

IV. Peptide Binding Affinity Datasets

A peptide is a short amino acid chain that is linked together by a peptide bond. Peptide binding has a critical importance in regulation of cellular signalling and understanding the mechanisms of the interactions between proteins and peptides. As there are many thousands of peptides, there is a need for prediction methods to help determine binding affinities of these peptides. In addition, in order to avoid this time consuming task, a computational predictive model can be developed. The difficulty of the peptide prediction problems when building a prediction model is the number of features being very large (around 6000) whereas the number of peptides in the training dataset is relatively small (≤ 150).

To test the proposed model, the high-dimensional peptide datasets provided at the Comparative Evaluation of Prediction Algorithms (CoEPrA) modeling competition [13] were used in this study and summarised in Table I and Table II.

As shown in Table I each task contains calibration (training) and prediction (test) datasets and physico-chemical descriptors have been provided for each small peptide in the datasets (for both calibration and prediction datasets). Each amino acid in a peptide is described by 643 descriptors. Task 2 consists of octa-peptides that have a total of 5144 descriptors. All other tasks have nona-peptides that have a total of 5787 descriptors. The task is to predict actual affinity values for peptides from the amino acid descriptors. The statistics (range, mean and standard deviation) of the binding affinities of the peptides of each task are given in Table II.

It is worth noting that only the first three tasks provided in CoEPrA modeling competition are related to the peptide binding problem, therefore were taken into consideration in this study, aim of which is to develop a robust predictive model for the prediction of binding affinity of the peptides.

V. Performance Measurements of the Prediction Models

There are different measurements used to assess capability of the predictive models. However, in order to maintain consistency over the published results and perform consistent comparison, coefficient of determination (q^2) has been used and is expressed as:

$$q^2 = 1 - \frac{\sum_{i=1}^{n}(y_{exp} - y_{prd})^2}{\sum_{i=1}^{n}(y_{exp} - \overline{y}_{exp})^2} \qquad (12)$$

where y_{exp} and y_{prd} are the expected and predicted values of the peptide, respectively, n is the number of peptides and $\overline{y}_{exp}$ is the mean of all expected values in the prediction dataset. The measure q^2 is a statistical model based upon the proportion of variability in a dataset [14]. When q^2 is close to 1 it suggests a model that has been successfully constructed. Negative q^2 values indicate that model poorly approximates the expected values.

VI. Results and Discussion

Prior to the analysis, the high dimensional datasets are normalized so that every feature fall within the same range of values. Then, the analysis is started by a feature selection method, namely MCFS method, which is used to reduce the dimension by estimating the quality of attributes in the data which then resulted in a low-dimensional physico-chemical attributes. Then TS fuzzy system with only two rules was constructed by using the reduced feature set. By using this rule-base the proposed method is able to build a robust and interpretable fuzzy system for a high-dimensional dataset with a relatively small number of data samples. FCM was used to identify the number of rules as well as the membership function parameters of the premise part whereas the coefficients of consequent part was determined by the SVR. Thus, the structure of the TS fuzzy system consitituted by the antecedent and consequent parts was determined.

The proposed model (TSFS-SVR) was applied to three tasks. Compared to the results as shown in Table III, the results are comparatively better than the recent studies presented in [1] and [15] for Tasks 2 and 3. The predictive performance for Tasks 2 and 3 have been improved by 5.5% and 17.4%, respectively. The parameters were selected properly to avoid overfitting in SVR. To address to this problem the grid-search method was applied. This method is not only

TABLE III
PREDICTION RESULTS OF THE COEPRA COMPETITION TASKS

Methods	Task 1 q^2	Task 2 q^2	Task 3 q^2
Step-Wise L1 Regularization [1]	0.667	0.642	0.205
Step-Wise L1, L2 Regularization [1]	0.691	0.668	0.131
KPLS exponential [15]	0.691	0.590	0.219
SVR	0.576	0.707	0.258
TSFS-SVR	0.653	**0.707**	**0.265**

simple and reliable but also allows parallel computations to speed up the calculations. The optimal parameters that yielded the best performance are found to be $C = 0.53$, $\epsilon = 0.05$ for Task 1, $C = 1.63$, $\epsilon = 0.10$ for Task 2, and $C = 0.25$, $\epsilon = 1.10$ for Task 3. Task 1, Task 2 and Task 3 contained 175, 225, 125 features, respectively.

The outcomes of the experiments performed evidently highlighted the strengths of SVR. Although the model has been constructed using small sample size in that the nature of peptide data, the predictive capability of the model proved to be of good generalization ability and more robust against the outliers. However quantifying peptide binding affinites requires more precision. Further experiments were carried out by using SVR alone using the optimised values of C and ϵ parameters in order to compare with the proposed model and see how fuzziness would effect its performance. The SVR alone yields poorer results for Tasks 1 and 3 whereas it remains the same for Task 2. One of the advantages of using fuzzy systems is its ability of managing uncertainty that exists in the datasets. The results clearly suggest that the fuzziness has positively contributed towards the modeling of the tasks. The results also appear to suggest that different sets of variables affect the result, and that exploration of the feature selection methods may further help accelerate the predictive power of the proposed hybrid method.

One difficulty for the analysis of post-genome data is the curse of dimensionality. The high-dimensional nature of this data negatively effects the performance of the prediction methods. Since thousands of features are available for peptides, a feature selection process was applied as an initial step to obtain low dimensional feature-set as the inputs of the fuzzy system. The final and best TSFS-SVR models are found to contain 175, 225 and 125 descriptors for Tasks 1, 2 and 3, respectively.

The rule-base of the proposed model contained only two rules which is the simplest form of a FS but resulted in better model. Therefore, there is more room for improvement by exploring different number of rules. Due to the page limitation, the rule-base driven as a result of TSFS-SVR will be presented and discussed during the presentation of the paper.

VII. CONCLUSIONS

In this paper, a hybrid system that has helped considerably improve the predictive capability of TSFS with the aid of SVR was developed and presented with the successful applications in the prediction of peptide binding affinity being regarded as one of the difficult modelling problems in bioinformatics.

The SVR-based experiments were carried out for three different peptide affinity datasets. The results suggest that the proposed hybrid method yields comperatively better results. In addition, the proposed system yields the interpretable rule-base while other methods studied in the literature still remain black-box.

In order to further deal with fuzziness and uncertainty this SV regression-based approach will be also explored and implemented for type-2 fuzzy system [16] in order to see if their predictive capability can be further improved. In addition, other optimisation methods including regularisation techniques will be studied for both type-1 and type-2 fuzzy systems. Given the promising results, different sequence-driven features and feature selection methods will be further explored along with varying number of rules.

REFERENCES

[1] O. Demir-Kavuk, M. Kamada, T. Akutsu and E. Knapp, "Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features," BMC Bioinformatics, vol. 12, pp. 412, 2011.

[2] L. A. Zadeh, "Fuzzy sets," Information and Control, vol. 8, pp. 338-353, 6, 1965.

[3] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," Systems, Man and Cybernetics, IEEE Transactions on, vol. SMC-15, pp. 116-132, 1985.

[4] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola and V. N. Vapnik, "Support Vector Regression Machines", in Advances in Neural Information Processing Systems 9, NIPS 1996, pp. 155-161, MIT Press, 1997.

[5] Jung-Hsien Chiang and Pei-Yi Hao, "Support vector learning mechanism for fuzzy rule-based modeling: a new approach," Fuzzy Systems, IEEE Transactions on, vol. 12, pp. 1-12, 2004.

[6] Chia-Feng Juang and Cheng-Da Hsieh, "A Fuzzy System Constructed by Rule Generation and Iterative Linear SVR for Antecedent and Consequent Parameter Optimization," Fuzzy Systems, IEEE Transactions on, vol. 20, pp. 372-384, 2012.

[7] R. Nikhil, Kuhu Pal, J. C. Bezdek and T. A. Runkler, "Some issues in system identification using clustering," in International Conference on Neural Networks, 1997, pp. 2524-2529.

[8] V. N. Vapnik, "An overview of statistical learning theory," Neural Networks, IEEE Transactions on, vol. 10, pp. 988-999, 1999.

[9] J. -. R. Jang, "ANFIS: adaptive-network-based fuzzy inference system," Systems, Man and Cybernetics, IEEE Transactions on, vol. 23, pp. 665-685, 1993.

[10] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology, vol. 2, pp. 1-27, 2011.

[11] J. C. Bezdek, "FCM: The fuzzy c-means clustering algorithm," Computers and Geosciences, vol. 10, pp. 191-203, 1984.

[12] D. Cai, C. Zhang and X. He, "Unsupervised feature selection for multi-cluster data," in 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 333-342.

[13] O. Ivanciuc, CoEPrA [http://www.coepra.org/], 2006, Last accessed 31/01/2013.

[14] R. G. D. Steel, J. H. Torrie, Principles and Procedures of Statistics. McGraw-Hill, 1960.

[15] C. Bergeron, T. Hepburn, C. M. Sundling, M. P. Krein, W. P. Katt, N. Sukumar, C. M. Breneman, K. P. Bennett, "Prediction of peptide bonding affinity: kernel methods for nonlinear modeling," CoRR abs/1108.5397, 2011.

[16] J. M. Mendel and R. I. B. John, "Type-2 fuzzy sets made simple," Fuzzy Systems, IEEE Transactions on, vol. 10, pp. 117-127, 2002.