Contemporary Engineering Sciences, Vol. 9, 2016, no. 16, 791 - 797
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2016.66114

Integrating Data Selection and Extreme Learning Machine to Predict Protein-Ligand Binding Site

Umi Mahdiyah

Universitas Nusantara PGRI Kediri, Mojoroto, Kediri, 64112, Indonesia

Elly Matul Imah

Universitas Negeri Surabaya, Ketintang, Surabaya, 50231, Indonesia

M. Isa Irawan

Institut Teknologi Sepuluh Nopember, Sukolilo, Surabaya, 60111, Indonesia

Abstract

Recently, computer-aided drug design is developing rapidly. The first step of computer-aided drug design is to find a protein - ligand binding site, which is a pocket or cleft on the surface of the protein being used to bind a ligand (drug). In this study, the binding site is defined as a binary classification problem to differ the location which can bind or cannot bind the ligand. Classification method used in this research is Extreme Learning Machine (ELM), because this method has fast learning process. In the real case, the dataset usually has imbalanced data. One of them is to predict binding site. Imbalanced data can be solved in several ways. In this study we carried out the integration of data selection and classification to overcome the inconsistency problem. The performance of integrating between data selection and Extreme Learning Machine to predict protein-ligand binding site is measured by using recall, specificity, G-mean and CPU time. The average of recall, specificity, G-mean and CPU time in this research are respectively, those are 91.8472%, 97.071%, 94.2647%, and 2.79 second.

Keywords: Extreme Learning Machine, imbalanced data, binding site protein-ligand

1 Introduction

Bioinformatics is a multidisciplinary science that involves many discipline science, such as molecular biology, mathematics, computational science, molecular chemistry, physics, and several other discipline science [1]. Actually, bioinformatics can be widely applied in various problems, such as to identify the host of SARS epidemics [2], and to gain drug design. The drug design concept on the bioinformatics based on the functionality of the protein.

Actually, there were many computation approaches that based on structure and sequence that have been developed to predict the binding site [3-9]. The prediction of the binding site can be formulated as a binary classification problem, that differ the location of the binding sites and non-binding site. Extreme Learning Machine (ELM) is an algorithm for pattern recognition and classification with a good performance [10-11]. ELM has been relatively computation faster than other neural networks. In addition according to [12] ELM has great accuracy and it is almost the same as Support Vector Machine (SVM) for balanced data.

As other bioinformatics data, protein-ligand binding site data has the imbalanced character. Imbalanced data is a considerable problem on machine learning, because it influent the performance of machine learning. There are many ways to overcome the problem of imbalanced data, one of them is undersampling. Undersampling is how to solve the problem of imbalanced data by reducing the majority data, so we can obtain the right proportion data and even the balanced data. According to Imah [13], a pattern recognition system has disadvantages like the condition of inconsistencies between data selection and classification while both steps are carried out separately, it is necessary to do the integration of the two steps.

Based on information above, in this study we use integrating data selection and ELM. The purpose of data seection is that it needs no big-size memory and long computation time. The integration is to overcome the problem of inconsistencies between data selection and classification process, because they are in the system [14].

2 Predicting Protein-Ligand Binding Site

LIGSITE is a geometry-based method to find a binding site [6]. An improvement of the LIGSITE algorithm developed by Levitt and Banaszak namely POCKET program. The program begins with a regular Cartesian grid. Secondly, the examination applied to the grid spacing to ensure protein atoms do not overlap with grid points. All grid points, which do not overlap with protein atoms, labeled as a solvent. If the grid points outside of the protein that is covered by surface proteins, that is grid points flanked by a pair of atoms in the protein.

It is called protein-solvent-protein (PSP) event. All residues in the protein are not necessarily always important. Some things are necessary for a vital protein structure and function of proteins, whereas others can be substituted. Analysis conservation is one of many methods used to predict functionally important residues of the protein sequence [5].

Table 1 Dataset of Protein

Number	Protein	Protein ID	Size of Data
1	Hydrolase	4TPI	3042
2		2ZAL	5286
3		2V8L	2060
4		1WYW	9616
5		1RN8	2233
6		1C1P	4797
7		1YBU	5909
8	Oxidoreductase	3D4P	6204
9		1A4U	4663
10		2WLA	2444
11	Transferase	2GGA	4146
12		1SQF	4365
13		1O26	9272
14		1G6C	4504
15		1BJ4	4205
16	Ligase	1U7Z	6144
17		1ADE	9072

3 Extreme Learning Machine (ELM)

ELM algorithm is derived from the minimum norm least square solution SLFNs. The main concept of the ELM as presented in the paper Huang [11] is as follows, given a training set $\aleph = \{(x_j, t_j) | x_j \in \boldsymbol{R}^{\boldsymbol{n}\times\boldsymbol{m}}, t_j \in \boldsymbol{R}^{\boldsymbol{n}},\ j \in [1, N]\}$, activation function $g(x)$, and hidden node number $\widetilde{N}$.

Step 1: Randomly assign input weight w_i and bias b_i, i=1,…, $\widetilde{N}$

Step 2 : Calculate the hidden layer output matrix H.

Step 3 : Calculate the output weight

$$\beta = H^{\dagger}T \tag{1}$$

where, $H = \begin{bmatrix} g(w_1.x_1 + b_1) & \cdots & g(w_{\tilde{N}}.x_1 + b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ g(w_1.x_N + b_1) & \cdots & g(w_{\tilde{N}}.x_N + b_{\tilde{N}}) \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{\widetilde{N}} \end{bmatrix}$, and $T = \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix}$

According to Liang [15], training process is done sequentially can affect the weight of update process. So that the weight of the output [10].

$$\beta_{n+1} = \beta_n + P_{n+1} H_{n+1}^T \left(T_{n+1} - H_{n+1} \beta_n \right) \tag{2}$$

where $P_{n+1} = P_n - P_n H_{n+1}^T (I + H_{n+1} P_n H_{n+1}^T)^{-1} H_{n+1} P_n$ (3)

$$P_0 = \left(H_0{}^T H_0 \right)^{-1}. \tag{4}$$

where β_{n+1} is β for $(n+1)$ data, β_n is β for n data, H is hidden layer matrix

4 Results and Discussion

4.1 Experimental Setting and Dataset

The proteins data are used to predict protein-ligand binding site can be seen in Table 1. In this study, the used data are experimental data proteins. That published in the RCSB Protein Data Bank web, it is also an open source data. We use the 17 proteins data, then we compare IDELM with ELM, BP (Backpropagation), and SVM (Support Vector Machine). The training and the testing process are done in every kids of protein, one protein for testing and the other one's for training. So the training process is done as many as the rest of data in each kind of proteins.

4.2 Experimental Result

Recall is the portion of the data samples correctly predicted by the algorithm. Then, specificity is accuracy of negative sample. From Table 2, we can see that the integration of data selection and classification recall is better than ELM recall in the issue of imbalanced data. The average of recall for IDELM is 0.918472, mean while the average value for a recall on a regular ELM is only 0.2944. It means 91.8472% of the data can be recognized correctly by IDELM. The average of recall from SVM and BP better than IDELM, these are 95.105% and 93.311%. Then, from this table also can be seen the best average of specivicity is IDELM, it is 0,97071

Table 2 recall and specificity of predicting protein-ligand binding site

Protein ID	**Recall**				**Specificity**			
	IDELM	**SVM**	**ELM**	**BP**	**IDELM**	**SVM**	**ELM**	**BP**
4TPI	0.963	0.982	0.222	0.989	0.981	0.613	0.795	0.732
2ZAL	0.918	0.961	0.484	0.973	0.999	1.000	0.500	0.998
2V8L	0.960	0.980	0.120	0.983	0.985	0.958	0.936	0.978
1WYW	0.870	0.989	0.000	0.990	0.923	0.399	1.000	0.292
1RN8	0.970	0.972	0.241	0.981	0.981	0.981	0.787	0.978
1C1P	0.930	0.961	0.622	0.939	0.990	0.984	0.408	0.971
1YBU	0.919	0.964	0.166	0.976	0.934	0.939	0.837	0.897
3D4P	0.957	0.872	0.500	0.932	0.999	1.000	0.500	1.000

Table 2 (Continued): Accuracy and recall of predicting protein-ligand binding site

1A4U	0.864	0.876	0.171	0.899	0.990	0.974	0.507	0.983
2WLA	0.842	0.977	0.500	0.972	0.997	1.000	0.500	1.000
2GGA	0.981	0.968	0.250	0.966	0.998	0.991	0.750	0.999
1SQF	0.931	0.813	0.520	0.981	0.995	0.977	0.595	0.990
1O26	0.943	0.935	0.361	0.980	0.870	0.782	0.573	0.604
1G6C	0.939	0.966	0.750	0.972	1.000	0.998	0.250	0.996
1BJ4	0.925	0.813	0.099	0.941	0.856	0.545	0.903	0.770
1U7Z	0.761	0.975	0.000	0.979	1.000	1.000	1.000	1.000
1ADE	0.942	0.859	0.000	0.721	1.000	0.931	1.000	0.083

G-mean and CPU time of IDELM, ELM, Backpropagation, and SVM for predicting protein-ligand binding site are shown in the Table 3. G-mean (Geometric mean) is a performance measurement tool for the evaluation of the imbalance data characteristics. From Table 3 can be seen that the G-mean average of ELM, IDELM, BP and SVM from 17 the data are 0.3546, 0.9426, 0.8777, and 0.9032.

From Table 3 also can be seen that the ELM is the algorithm has a fast CPU Time. In this study, the fastest average of CPU Time is ELM, it is 0.069283 seconds. Whereas the most CPU time is at BP, it is 22.18663 seconds, while CPU time of IDELM and SVM are 2.786071 and 2.4443 seconds.

Table 3 G-mean and CPU Time of predicting protein-ligand binding site

Protein ID	***G-mean***				***CPU Time*(s)**			
	IDELM	**SVM**	**ELM**	**BP**	**IDELM**	**SVM**	**ELM**	**BP**
4TPI	0.972	0.690	0.084	0.804	2.475	1.076	0.039	11.019
2ZAL	0.957	0.980	0.000	0.985	2.506	1.277	0.044	13.268
2V8L	0.972	0.967	0.111	0.980	2.774	1.113	0.023	10.798
1WYW	0.895	0.474	0.000	0.361	1.986	1.261	0.070	23.158
1RN8	0.975	0.976	0.094	0.980	2.691	1.105	0.016	13.632
1C1P	0.959	0.972	0.095	0.953	2.582	1.082	0.036	16.352
1YBU	0.924	0.948	0.028	0.934	2.431	0.991	0.042	7.592
3D4P	0.978	0.933	0.000	0.965	1.989	5.866	0.226	16.918
1A4U	0.924	0.923	0.034	0.939	2.395	1.209	0.062	11.411
2WLA	0.916	0.988	0.000	0.986	2.613	1.295	0.047	59.780
2GGA	0.989	0.979	0.000	0.982	2.441	1.907	0.062	12.059
1SQF	0.962	0.882	0.221	0.986	3.050	1.342	0.042	15.002
1O26	0.905	0.852	0.119	0.663	2.363	1.845	0.101	13.135
1G6C	0.969	0.982	0.000	0.984	3.015	1.888	0.055	23.665

Table 3 (Continued): G-mean and CPU Time of predicting protein-ligand binding site

1BJ4	0.886	0.637	0.198	0.841	3.908	1.903	0.047	25.019
1U7Z	0.872	0.988	0.000	0.990	4.852	1.794	0.062	9.142
1ADE	0.970	0.895	0.000	0.245	3.292	14.601	0.203	75.223

5 Conclusions

ELM has a good average of CPU time in almost all the data is 0.006348 second. CPU Time of SVM, IDELM, and BP are 0.078306, 0.0637, and 0.169536 second respectively. From the result of some methods, IDELM has the best average recall, specificity, and G-mean. Recall is 97.0471%, specificity is 97.071%, and G-mean is 94.2647%..

References

[1] S. Shen, and J.A. Tuszynski, *Theory and Mathematical Methods for Bioformatics,* Springer, Verlag Berlin Heidelberg, 2008. http://dx.doi.org/10.1007/978-3-540-74891-5

[2] M. Isa Irawan and S. Amiroch, Construction of Phylogenetic Tree Using Neighbor Joining Algorithms to Identify the Host and Spreading of SARS Epidemic, *Journal of Theoretical and Applied Information Technology (JATIT),* **71** (2015), 424- 429.

[3] G.Y. Wong, F.H.F. Leung and S.H. Ling, Predicting Protein-Ligand Binding site Using Support Vector Machine with Protein Properties, *Transactions on Computational Biology and Bioinformatics*, **10** (2013), 1517-1529. http://dx.doi.org/10.1109/tcbb.2013.126

[4] G.Y. Wong, F.H.F. Leung and S.H. Ling, Predicting Protein-Ligand *Binding site* with Differential Evolution and Support Vector Machine, *The 2012 International Joint Conference on Neural Networks (IJCNN)*, Brisbane, Australia, 2012. http://dx.doi.org/10.1109/ijcnn.2012.6252744

[5] J.A. Capra and M. Singh, Predicting functionally important residues from *sequence* conservation, *Bioinformatics*, **23** (2007), 1875-1882. http://dx.doi.org/10.1093/bioinformatics/btm270

[6] M. Hendlich, F. Rippmann, and G. Barnickel, LIGSITE: Automatic and Efficient Detection of Potential Small Molecule-Binding Sites in Proteins, Journal of Molecular Graphics and Modelling, 15 (1997), 359-363. http://dx.doi.org/10.1016/s1093-3263(98)00002-3

[7] Roman A. Laskowski, SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions, *Journal of Molecular Graphics*, **13** (1995), 323-330 http://dx.doi.org/10.1016/0263-7855(95)00073-9

[8] A. Gutteridge, G. Bartlett and J. Thornton, Using a Neural Network and Spatial Clustering to Predict the Location of Active Sites in Enzymes, *J. Molecular Biology*, **330** (2003), 719-734. http://dx.doi.org/10.1016/s0022-2836(03)00515-1

[9] A. Koike and T. Takagi, Prediction of Protein-Protein Interaction Sites Using Support Vector Machines, *Protein Engginering Design and Selection*, **17** (2004), 165-173. http://dx.doi.org/10.1093/protein/gzh020

[10] Umi Mahdiyah, M.I Irawan, E.M Imah, Study Comparison Backpropogation, Support Vector Machine, and Extreme Learning Machine for Bioinformatics Data. *Journal of Computer Sciences and Information*, **8** (2015), 55-62.

[11] G. Huang, Q. Zhu and C. Siew, Extreme Learning Machine: Theory and Applications, *Neurocomputing*, **70** (2006), 489-501. http://dx.doi.org/10.1016/j.neucom.2005.12.126

[12] J. Chorowski, J. Wang, and J. M. Zurada, Review and Performance Comparison of SVM- and ELM-based Classifiers, *Neurocomputing*, **128** (2014), 507-516. http://dx.doi.org/10.1016/j.neucom.2013.08.009

[13] E.M. Imah, W. Jatmiko, and T. Basarudin, Adaptive Multilayer Generalized Learning Vector Quantization (AMGLVQ) as new algorithm with integrating feature extraction and classification for Arrhythmia heartbeats classification, *IEEE International Conference Systems, Man, and Cybernetics (SMC)*, (2012), 150-155. http://dx.doi.org/10.1109/icsmc.2012.6377692

[14] Umi Mahdiyah, M. Isa Irawan, E.M. Imah, Integrating Data Selection and Extreme Learning Machine for Imbalanced Data, *Procedia Computer Science*, **59** (2015), 221-229. http://dx.doi.org/10.1016/j.procs.2015.07.561

[15] N. Liang, G. Huang, P. Saratchandran and N. Sundararajan, A Fast and Accurate Online Sequential Learning Algorithm for Feedforward Networks, *IEEE Transactions on Neural Networks*, **17** (2006), 1411-1423. http://dx.doi.org/10.1109/tnn.2006.880583

Received: November 15, 2015; Published: July 27, 2016