International Journal of Database Theory and Application
Vol.8, No.5 (2015), pp.207-214
http://dx.doi.org/10.14257/ijdta.2015.8.5.17

How to Address the Problems of Data Mining in Chemistry: The Application of Artificial Neural Networks

Boqin Liu

Faculty of Computer and Information Science, Southwest University, China
boqliu@swu.edu.cn

Abstract

Recent years, data mining has become a very popular concept in computer science. Its application has covered many crucial subjects. However, the systemic research based on relevant techniques in the field of chemistry is still non-existent. Here, we present a future possibility of the osmosis of data mining to chemistry under the circumstance of the age of big data, using artificial neural network (ANN) models as a crucial example. By presenting its applications in different research areas, this paper gives a comprehensive understanding to the ANN and its potential to dominate the chemical data mining area.

Keywords: *Data mining; big data; artificial neural network; chemistry*

1. Introduction

Data mining is an interdisciplinary subfield of computer science, which is the computational process in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems [1-3]. The ultimate goal of the data mining process is to extract useful information from a database and transfer it into an acceptable structure for further use [1]. In addition to the initial analytical stage, it contains data sets and data managements, interestingness metrics, data processing and pre-processing, modeling and artificial considerations, complexity considerations, and post-processing. The concept has become extremely popular among scientific research.

The ultimate task of data mining is the supervised or unsupervised analysis of large scale of data, in order to extract the unknown interesting patterns. It often contains the usage of database techniques. These patterns can then be regarded as a kind of sum of the input data, and can be utilized in further analysis. In scientific research, one of the most frequently-used techniques is the machine learning techniques. Machine learning is a scientific discipline that finds out the construction and the exploration of algorithms that can learn from existing data [4-6]. These algorithms are based on inputs and can be used for making predictions or decisions, but they are not following only explicitly programmed instructions. In the research on chemistry, artificial neural network (ANN) approaches is one of the most usual tool to address the problem of the large scale data. Zupan and Gasteiger [7] was the first scientists to put forward a concluded concept of the application of ANN models to chemistry. Since then, thousands of studies using ANN techniques as a supplement or as the main objective in the chemical research were reported. In the next section, we aim at introducing the principle of ANN models and listing some of the representative kinds of neural networks which are frequently-used in chemical researches.

2. Artificial Neural Network

Artificial neural network (ANN) is an information processing system with interconnected components analogous to neurons [8-13]. It is on the basis of the non-linear functions and the connection of the assumed "neurons" (which is what we called

ISSN: 2005-4270 IJDTA

"network"). On the basis of mathematical models, it can simulate some functions of the biological neural networks. ANN models can avoid the noise in the large scale of data, and generate the correct and robust responses with very low RMS errors.

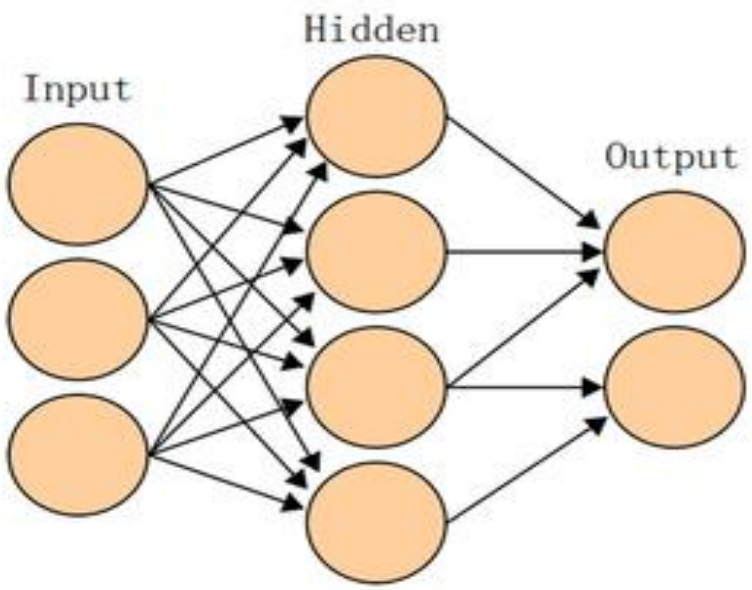

Figure 1. General Structure of an Artificial Neural Network (Two Layers, Three Input Variables, Four Nodes and Two Output Variables)

Figure 1 shows the main structure of an ANN model with two layers, three input variables, four nodes and two output variables. A completed ANN model consists of three parts: the input layer, the output layer and the hidden layer. The input layer is made up of independent variables. Usually, we don't consider the input layer as one of a layer in an artificial neural network. Hidden layer is a layer existing between the input and output layer. The more the number of the neurons in the hidden layer has, the greater robustness of the artificial neural network will be. After a series of computing processes, the output variables are exported from the output layer.

Back-propagation neural network (BPNN) is the most frequently-used ANN models in chemical research [14-16]. More than 80% ANN-chemical research used BPNN models to development their prediction and pattern recognition models [7]. The widely use of BPNN model is based on its strong adjustment ability of the weights. According to the algorithms of BPNN model, the errors of the calculations can be converted back to the last layer and can be revised if it is not precise enough. The main framework of this function of BPNN model is shown in Figure 2:

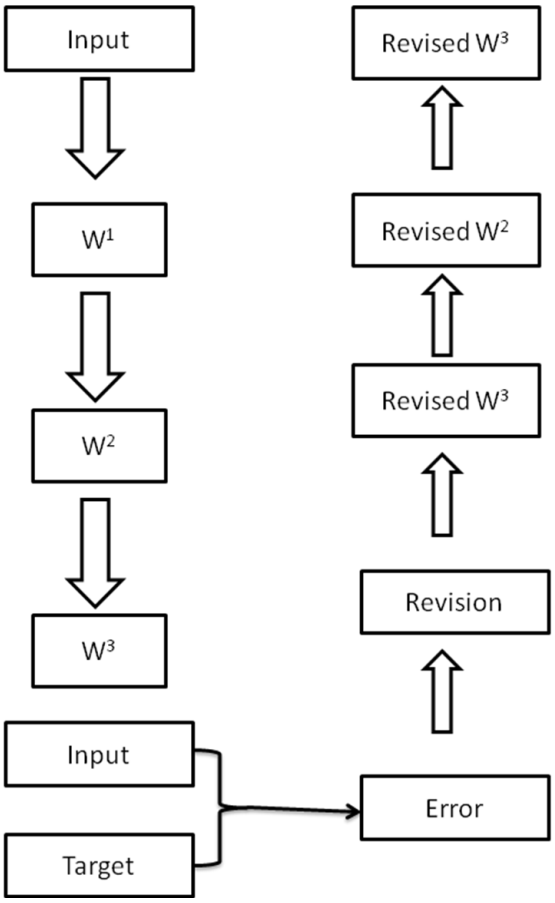

Figure 2. Weight Revision Principle of BPNN

Where W is the weight for the calculation. The strong self-adjustment of the weights and errors ensure the robustness and correctness of the BPNN model. Besides, with the development of the SPSS software, BPNN model has become extremely popular among relevant scientific research. Scientists can choose different number of the layers and nodes to find out the best structure of the BPNN model, based on the RMS errors.

In addition to the BP neural network, there are many other neural networks that were used for addressing the chemical problems. For example, single-layer neural networks like Hopfield network and Kohonen network also are of great importance in early chemical and chemical related researches. Recent years, as the development of many packed and user-friendly software, general regression neural network (GRNN) [17, 18] and multi-layer feed forward neural network (MLFN) [19-20] have also become very popular in the studies of chemistry, materials and environment. Especially, the GRNN model is highly effective because of its high prediction function and the development of relevant software ensure its highly robust maneuverability. In the future, with the quick developments of the package of various ANN models, there will be more commercial software that promotes the use of advanced ANN models.

3. Chemical Prediction Models Using ANN Techniques

As for the practical application of ANN models in chemistry, here, we present some of the preventative previous researches about the ANN models in the field of chemistry.

Howard and Karplus [21] used a special recognition approach to recognize the structure and sequence of proteins. Here we use a typical example to illustrate this easy but useful approach: Figure 3 shows the sequence of the protein in the amino acid sequence of ASN-TYR-TYR-ALA-MET-ALA-MET-MET:

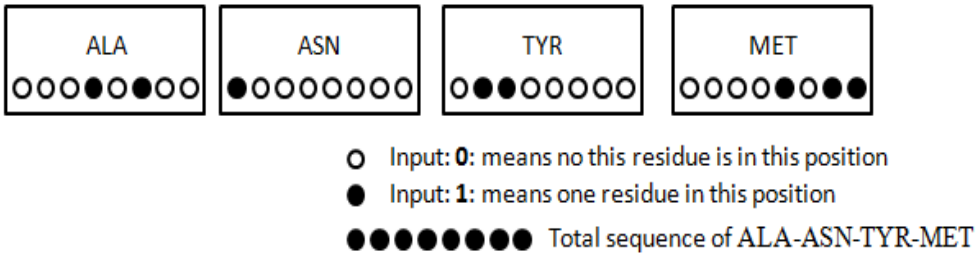

Figure 3. An Example of the Protein Recognition Approach for ANN Non-Linear Learning

Using this recognition method, we can only input "0" or "1" to present the sequence of a single protein chain. The groups of inputs and the number of the circles can be extended in order to adjust in accordance with the practical applications. By using this recognition approach, one can develop various models for the prediction of many crucial characteristics of proteins such as the secondary structure and other crucial properties which are not easy for scientists to obtain from practical experiments and industries.

Besides, there are many other excellent researches which used ANN models for the prediction of chemical properties. Valderrama and colleagues [22] predicted the ensity of ionic liquids using group contribution and artificial neural networks. Albahri and George [23] used ANN models to predict the auto ignition temperature (AIT) of pure components based on group contribution principles. Hao and his colleagues [24] used a novel and convenient recognition approach for the hydrocarbon compounds, and successfully predict the permittivities and polarizabilities using GRNN and MLFN models with a series of comparative low RMS errors.

All the representative researches above are based on a relatively large scale of data. And results confirm that ANN models can generate correct responses when the data scale is large enough. Based on these researches, we strongly believe that in the field of data mining, ANN models can be used for the prediction of chemical properties from the large scale of data. The lager the scale of the scale is, the more correct of the ANN results will be. Obviously, the data scale in these research [21-24] is not very high, which do not correspond with the real concept of the big data. But we can assume that the results of the ANN models will be much more precise with larger scales under the circumstance of big data. Being combined with the most advanced techniques such as cloud computing. ANN models will definitely be promoted to a higher stage.

4. Chemical Pattern Recognition Models Using ANN Techniques

In the field of chemistry and chemical related researches, there also many outstanding researches that are based on ANN approaches [25-26].

Compared to prediction models based on ANN approaches, pattern recognition seems more difficult to generate the correct results. The main reason is that ANN model, as a non-linear machine, it usually generates over-fitting results in the applications of pattern recognition. Since the development of support vector machine (SVM) [27-29], ANN pattern recognition models are gradually being replaced by SVM techniques. However, SVM techniques cannot thoroughly replace ANN because SVM can only recognize two different objects, which cannot be extended to a wider application. The reason why the ANN models are now being replaced is that pattern recognition is that ANN models usually generate the over-fitting results [30-31] when training limited data. Only a very large scale of data can we ensure the robustness and correctness of ANN models.

According to the concept of big data and data mining, the utility value of the ANN models for pattern recognition may be reinvented under a large scale of data. That is to say, there is a renascent tendency that ANN models can be widely researched again using a large scale of data with the development of cloud computing and advanced database techniques.

5. A Case Study Using ANN Models

To highlight the powerful function of ANN models in chemistry and chemical related disciplines under the background of big data, we used a complex example which is not easy to be described by linear functions or other techniques in order to illustrate the adaptability of ANN models and relevant non-linear functions. Here, we use GRNN and MLFN models to repeat one of the previous studies. Because the results of ANN models are mainly decided by the initial stochastic value of the weights and the components of training and testing sets, results can be different after each repeated training and testing process. We used the data from reference [32], analyzing and predicting the non-linear relationship between different ions and pH values. We used the typical expression of the result table and figures provided by reference [24] so that the results would be very clear and intuitionistic.

Table 1. Results of Models for the Analysis of Rainwater

ANN Model	Trained Samples	Tested Samples	Average RMS Error
Linear Regression	89	22	2.77
GRNN	89	22	0.13
MLFN with 2 Nodes	89	22	1.87
MLFN with 3 Nodes	89	22	1.41
MLFN with 4 Nodes	89	22	1.56
MLFN with 5 Nodes	89	22	1.99
MLFN with 6 Nodes	89	22	3.24
MLFN with 7 Nodes	89	22	4.14
MLFN with 8 Nodes	89	22	5.21
MLFN with 9 Nodes	89	22	4.78
MLFN with 10 Nodes	89	22	7.21
MLFN with 11 Nodes	89	22	7.89
MLFN with 12 Nodes	89	22	9.13
MLFN with 13 Nodes	89	22	10.24
MLFN with 14 Nodes	89	22	11.77
MLFN with 15 Nodes	89	22	21.13
MLFN with 16 Nodes	89	22	23.39
MLFN with 17 Nodes	89	22	30.57
MLFN with 18 Nodes	89	22	28.61

Table 1 show that the GRNN model can generate an extremely good result, with a very low RMS error. Results reveal that the GRNN model can usefully predict the pH values of rainwater. Here we present two groups of figure to highlight the training and testing results of the GRNN model.

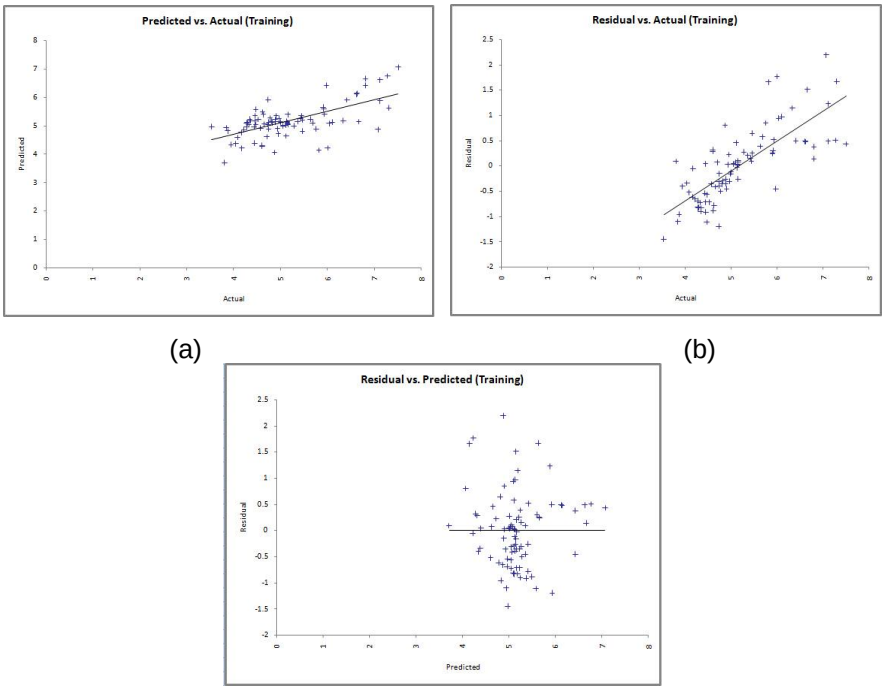

Figure 4. Training Results for the Analysis of Rainwater Using GRNN Model

Figure 4 shows the training process (recall test) of GRNN model. Results show that using a relative big scale of data, GRNN model still can generate a very good response. Figure 4 (a) show that the predicted values are quite close to the actual values. And Figure 4 (b) and (c) show that the residual values are comparative low, closing to zero.

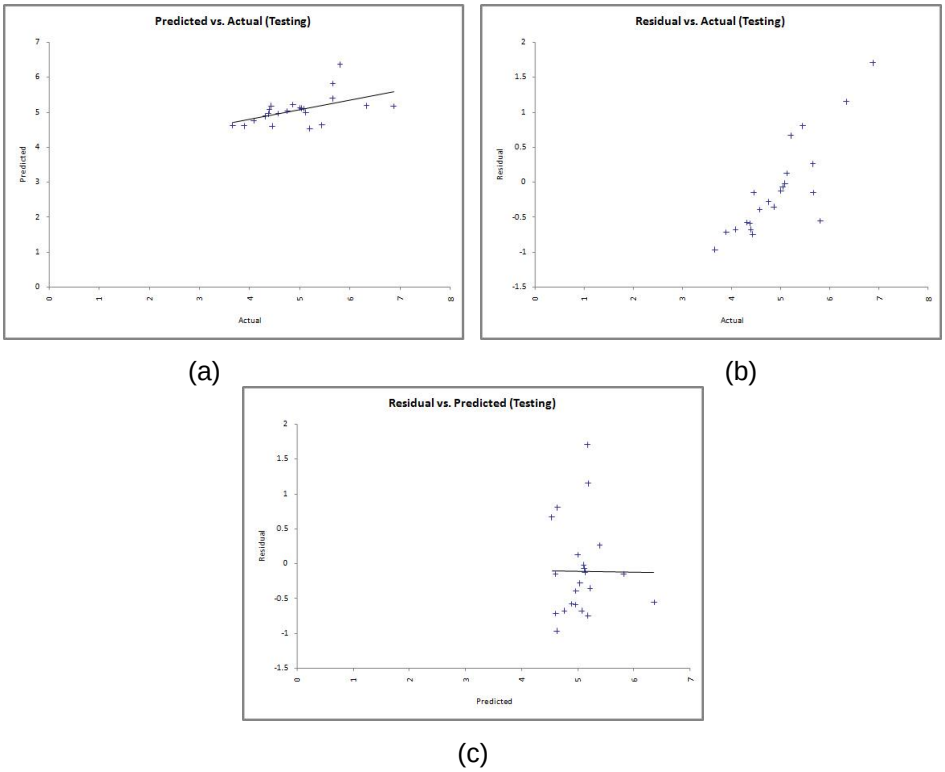

Figure 5. Testing Results for the Analysis of Rainwater Using GRNN Model

Figure 5 shows the testing results of the GRNN model. Figure 5 (a) shows that the predicted results are very close to the form of actual values, illustrating that the results are robust. Figure. 5 (b) and (c) shows that the residual values are low.

According to this case study, ANN model especially the GRNN can analyze and predict the pH values of rainwater with an extremely low RMS error. This example illustrate that using a large scale of data, ANN models have a great potential application to scientific research, corresponding with the concept and ultimate goal of big data and data mining.

ANN models are powerful tools that can give a very prefect result under a large scale of data. However, all the results are generated from the "black box". We do not know the exact form of the fitted results. The only thing we can obtain from the training and testing processes is the weight, errors and outputs, which lead to the limitations of ANN models. So far, few research can use ANN model to determine the exact presented non-linear relationship during the model development of ANN models. That is to say, ANN models lack the ability to find out the causal relationship between independent variables and dependent variables. This limitation also cause the bottleneck of relevant research. Nevertheless, we still can find out the very precise results from ANN models under the condition of big data. We're still confident about its prospect for development because it is still irreplaceable in related scientific research. That is why we believe that the age of data mining and big data can refresh the development of ANN models to the application in chemistry and chemical related disciplines.

6. Conclusion

In this perspective, we combine previous works of artificial neural networks (ANNs) models with relevant research on chemistry and chemical related researches. We show that ANN models are powerful tools that can be used for the predictions and pattern recognitions. Now all of us are undergoing the wave of data mining, and ANN models have successfully shown their great vitality in the large scale of data. Due to the limitation of the data scale, previous ANN-based chemical research couldn't meet the new requirements of the "big data century". But the robust and precise results shown by previous researches have proved that ANN models have very good responses to it. We can assume that under the circumstance of a larger scale of data, ANN models in the fields of chemistry, materials science and environmental science can be promoted to a higher stage in the next several years. In future studies, we will pay our attention to the development of ANN models and relevant algorithms that can suit for the large scale of data and apply these novel techniques to chemistry and chemical related researches.

Acknowledgment

This work was jointly supported by for the Education research project of Southwest University (NO.2013JY066), Fundamental Research Funds for the Central Universities (XDJK2014C143).

References

[1] N. Sharma and H. Om, “Network Modeling Analysis in Health Informatics and Bioinformatics”, vol. 2, no. 4, **(2013)**.
[2] A. V. Kate, P. V. Nikilav, S. Giriesh and J. Naren, “Multimedia Data Mining-A Survey”, **(2013)**.
[3] Y. B. Qin and D. X. Lu, “Applied Mechanics and Materials”, vol. 263, **(2013)**.
[4] C. Andrieu, N. De Freitas and A. Doucet, “Machine Learning”, vol. 50, no. 1-2, **(2003)**.
[5] D. E. Goldberg and J. H. Holland, “Machine Learning”, vol. 3, no. 2, **(1988)**.
[6] C. E. Rasmussen, “Gaussian processes for machine learning”, **(2006)**.
[7] J. Zupan, J. Jure and J. Gasteiger, “Anal. Chim. Act”, vol. 248, no. 1, **(1991)**.
[8] B. Samanta, K. R. Al-Balushi, “Mech. Syst. Signal Pr.”, vol. 17, no. 2, **(2003)**.
[9] W. G. Baxt, “Ann. Int. Med.”, vol. 115, no. 11, **(1991)**.

[10] M. Shao, X. J. Zhu and H. F. Cao, "Energy", vol. 67, **(2014)**.
[11] C. H. Aladag, A. Kayabasi and C. Gokceoglu, "Neural Computer App.", vol. 23, no. 2, **(2013)**.
[12] J. Van Schependom, G. Nagels and W. Yu, "Schizophrenia Bulletin", **(2013)**.
[13] E. C. Santos, E. D. Armas and D. Crowley, "Soil Bio. Bioche", vol. 69, **(2014)**.
[14] P. Wang, X. Ji and L. Zhu, "Sensors and Actuators A: Physical", vol. 201, **(2013)**.
[15] P. Wang, L. Zhu and Q. Zhu, "NDT & E International", vol. 55, **(2013)**.
[16] K. Z. Huang, X. K. Xiong and C. M. Zhang, "E Protein and peptide letters", **(2014)**.
[17] H. Yip, H. Fan and Y. Chiang, "Auto. Const.", vol. 38, **(2014)**.
[18] C. M. Hong, F. S. Cheng and C. H. Chen, "Int. J. Elect. Power & Energy System", vol. 60, **(2014)**.
[19] D. Svozil, V. Kvasnicka and J. Pospichal, "Chemo. Intel. Lab. System", vol. 39, no. 1, **(1997)**.
[20] T. D. Sanger, "Neural Networks", vol. 2, no. 6, **(1989)**.
[21] H. L. Howard and M. Karplus, "Proc. Nat. Aca. Sci.", vol. 86, no. 1, **(1989)**.
[22] O. Valderrama, J. A. Reátegu and R. E. Rojas, "Indust. & Eng. Chem. Res", vol. 48, no. 6, **(2009)**.
[23] A. Albahri, Tareq and R. S. George, "Indust. & Eng. Chem. Res", vol. 42, no. 22, **(2003)**.
[24] H. Li, X. F. Liu and S. J. Yang, "Int. J. Electro chem. Sci.", vol. 9, no. 7, **(2014)**.
[25] N. Bhat and T. J. McAvoy, "Computer & Chem. Eng.", vol. 14, no. 4, **(1990)**.
[26] J. C. Hoskins and D. M. Himmelblau, "Computer & Chem. Eng.", vol. 12, no. 9, **(1988)**.
[27] J. A. K. Suykens and J. Vandewalle, "Neural Proceeding Let", vol. 9, no. 3, **(1999)**.
[28] T. S. Furey, N. Cristianini and N. Duffy, "Bioinfor", vol. 16, no. 10, **(2000)**.
[29] G. Cauwenberghs and T. Poggio, "Adv. Neural Inf. Proc. System", **(2001)**.
[30] Z. H. Zhou, J. Wu and W. Tang, "Artificial intelligence", vol. 137, no. 1, **(2002)**.
[31] A. K. Jain, J. Mao and K. M. Mohiuddin, "Computer", vol. 29, no. 3, **(1996)**.
[32] G. Tang and N. Bai, "Acta Scientiea Circumstantiae", vol. 20, no. 5, **(2000)**.

Author

Boqin Liu, a graduate of Faculty of Computer and Information Science, Southwest University, Chongqing, China, in 2007 received master of computer education, work units is the Southwest University of Faculty of Computer and Information Science, associate professor, engaged in teaching and research university computer twenty years, the main research areas computer application technology, database applications and data mining.

