International Journal of Research in Advent Technology, Vol.3, No.9, September 2015
E-ISSN: 2321-9637
Available online at www.ijrat.org

Predicting Melting Point of Medicinal Compounds Using Neural Network Classifier

Surabhi P V[1], Dr. T.S Santha[2]
Research Scholar[1], Principal [2], Dr.GRD College of Science, Coimbatore, India[1, 2]
Email:surabhi.sreekumar@gmail.com[1], principal.cs@gmail.com[2]

Abstract- The determination of melting point range, chemical purity and solubility in the process of characterizing medicines is one of the principal requirements evaluated in quality control of the pharmaceutical industry. The ability to predict melting point of medicines is of great importance as it facilitates efficient lead candidate selection and drug development. The purpose of the present work is to develop a chemical expert system using Radial Basis Function Network (RBFN) for prediction of melting point of medicines. K- means clustering algorithm is used for the designing of Radial Basis Function Network and learned a linear regression on top of that. An algorithm is developed for feature selection. Six vital factors, those affecting the melting point of medicinal compounds are used as input for RBFN and output being the predicted melting point. A dataset of 100 medicines is used for system training and testing using WEKA workbench.

Index Terms- Artificial Neural Networks; Radial Basis Function Networks; Melting Point; Molecular Descriptors;QPSR/QSAR.

1. INTRODUCTION

1.1 Melting Point

Melting point is the temperature at which a compound’s solid phase is in equilibrium with its liquid phase. It is a fundamental physical property of organic compounds, which has found wide use in chemical identification, as a criterion of purity. Melting point is routinely used for determining physio chemical properties such as identity, purity and other properties of a substance like boiling point, water solubility and liquid viscosity [1]. Melting point affects solubility which controls toxicity. A poorly soluble compound’s concentration in the aqueous environment may be too low for it to exert a toxic effect [2]. The melting point of solid compounds mainly depend upon the molecular shape, intermolecular forces between molecules like London dispersion forces, hydrogen bonding forces and other intermolecular interactions [3]. Melting point is a descriptor for predicting the physiochemical properties of a compound [4]. Previously it was required to experimentally determine the melting point in laboratories for determining these properties. This may be hazardous, toxic, expensive and also may lead to the wastage of chemicals and time. There for a computational model for melting point prediction is highly essential and economical.

1.2 Data Mining

Data Mining is a database analytical process that search unknown patterns in data which can be used to predict future behavior. Data Mining is the task of automatic or semi automatic analysis of large set of data to extract interesting, unknown patterns through process like cluster analysis, anomaly detection and association rule mining. Extraction of knowledge in a human understandable structure from data is the main goal of data mining. The three stages of data mining process are

1.2.1 Data Exploration

Exploration phase starts with data preparation or data preprocessing which involve cleaning data, data transformations and feature selection operations to bring the number of variables to an adaptable range.

1.2.2 Model Building

In this stage various models are built and the best one is selected based on their predictive performance. A variety of techniques are developed to achieve the goal, many of which are based on Competitive Evaluation of Models- applying different models to same dataset and comparing their performance to select the best [5].

1.2.3 Deployment

Deployment stage involves using the best selected model and applying it to new data for creating predictions and approximations of the expected outcome. But the output of data mining depends on the dataset and the algorithm used. Sometimes the algorithm is not suitable for the given dataset so that the data is not classified as per need of the problem. The real task is to find the suitable algorithm for the given dataset.

2. LITERATURE REVIEW

Many researches have been conducted for predicting melting point of chemical compounds. Many methods like Artificial Neural Networks(ANN), K Nearest Neighbor (KNN), Principal Component Model, Support Vector Machine(SVM), Clustering, Random Forest etc. have been used in the past for this purpose. A model for predicting melting point of organic compounds based on their nature properties using Multilayer Perceptron Artificial Neural Network has been introduced by Nazarabadi, Taheri and Boshrooyeh [6]. This method uses a two layer artificial neural networks to model the nature of compounds and predicts the melting point of organic compounds. Rafidha, Balakrishnan and Sherly used a kernel based classification technique SVM Regression for the prediction of melting point of drug-like compounds in terms of topological descriptors, connectivity indices and 2D auto correlations [7]. Aziz, Pourbasheer and Danandeh applied Quantitative Structure Property Relationship to predict the melting point of drug-like compounds using the Principal Component - Genetic Algorithm – Multi parameter Linear Regression (PC-GA-MLR) and Principal Component – Genetic Algorithm – Artificial Neural Network (PC-GA-ANN) [8]. Results of both the methods were compared and the superiority of PC-GA-ANN over PC-GA-MLR has been demonstrated. Shodhganga investigated the predictability of three important Machine Learning Techniques for melting point prediction – KNN, ANN and SVM [9]. Better performance is achieved using SVM model than ANN and KNN. It is also found that performance was slightly better for hybrid of SVM and ANN than the performance of individual. Karthikeyan, Glen and Andreas were the first to develop a general model that covers a comparatively large and relevant part of organic chemical space using Feed-forward Back propagation Artificial Neural Network. It is based on a diverse dataset of 4173 compounds [10]. A comparison of KNN and K-star prediction models for melting points has been performed by Rafidha, Balakrishnan and Sherly [11]. It was determined that KNN provides more accuracy and less computation time than K-star.

3. PROPOSED WORK

Proposed work is based on Artificial Neural Network to predict the melting point of medicines more accurately. Artificial Neural Network is popular in QSPR/QSAR models where complex nonlinear relationships exist among data [13]. ANN is formed artificial neurons connected with weights, which are organized in layers. The neuron layers between input and output layers are called hidden layers. ANN does not need explicit formulation of the mathematical relationships of the problem stated. For these reasons ANN has been applied to a variety of chemical problems recently. Many machine learning models has been devised so far using data mining methods like Classification, Association and Clustering. But for majority of the models, the deviation of the result from the actual melting point is clearly distinguishable. There for a computational model which predicts melting point from a compound’s chemical properties more accurately with less computation time is highly desirable.

4. DATA AND METHODOLOGY

4.1 Dataset and Molecular Descriptors

Melting point dataset for 100 drug-like compounds were taken from the dataset compiled and published by Bergestorm [4]. Dataset include some liquid compounds but most of the compounds are solid at room temperature. The melting points are spread between 40˚ and 289˚ on the standard Celsius scale. Dataset contains compound’s experimentally calculated melting point and a SMILES (Simplified Molecular Input Line Entry System) description for calculating molecular descriptors. Molecular descriptors are the molecular properties which characterize the molecules in a compound. Molecular descriptors are calculated with the help of chemical software E-Dragon [12]. As a result, a total of 200 descriptors were calculated for each compound in the dataset (100 compounds).

4.2 Data Preprocess

Quality of data is a key factor for the success of any data mining algorithm. In this research, we developed many pre reduction steps to improve accuracy and computational time. Irrelevant and noisy data affects the accuracy of prediction. Irrelevant theoretical molecular descriptors are eliminated using the following procedures:

- Descriptors that are having constant values have been eliminated. We developed an algorithm for this purpose. A data matrix is designed with Molecular ID as rows and descriptors as columns. Algorithm scans entire matrix to determine columns having constant descriptor values. Such descriptors are eliminated from the dataset.
- Mean and Variance for each descriptor in the data matrix are calculated. Algorithm finds the descriptors that are having low variance and are removed from the dataset.
- Compounds which has melting point in between 120˚ and 190˚ Celsius are chosen.

We used C# and Microsoft .NET Framework as platform only as means for writing programming code [16].

It is crucial to employ appropriate descriptors to obtain a significant correlation. As the number of descriptors increases, the model becomes more complicated and the interpretation is difficult. For selecting the attributes, we used the attribute evaluator CfsSubsetEval and GreedyStepwise search method.

CfsSubsetEval is a flexible supervised attribute filter which can be used to select attributes. It allows various search and evaluation methods to be combined. It evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them [14].

GreedyStepwise search method performs a greedy forward or backward search through the space of attribute subsets. It starts with no/all attributes or from an arbitrary point in the space. It stops when the addition/deletion of any remaining attributes results in a decrease in evaluation. Finally six vital descriptors given in Table 1 were selected.

Table 1. Selected descriptors

Si. No.	Descriptor
1	DECC
2	X1A
3	X0AV
4	GATS6v
5	GATS7e
6	GATS5p

Algorithm for removing identical columns and Low variance descriptors.

Input: A dataset matrix A [m][n],m = 0:R , n = 0:C
Output: A dataset matrix A2 [m][r],m = 0:R, n = 0:P

```
1.  For all i such that $0 \le i \le C$ do : Label1
2.  Set temp=A[0][i]
3.  For all j such that $1 \le j \le R$ do
            If temp ≠ A[j][i] then
              Go to Label1
            End if
    End For
4.  If j=R-1 then
            Store column number in array D
            Mark the column as "deleted"
            Increment count
    End if
    End For
5.  Obtain the matrix A1[m][n – count] by
    removing columns in D from A
6.  Define Min_Variance
7.  For all i such that 0<i<(n-count)
8.  Set S = 0, M = 0, V = 0, K = 0, N=0, Error
    = 0
9.  For all j such that 0<j<R
10. Compute S = S + A1[j][i]
11.  If j = R-1 then
12.         Mean[M] = S/R
13.         Increment M
14. End if
15. End For
16. End For
17. For all i such that 0<i<(n - count)
18. For all j such that 0<j<R
19. Error = Error + $(A1[j][i] – Mean[K])^2$
20. If j=R-1 then
21.         Var[V] = Error/(R - 1)
22.         Increment V
23. End If
24. End For
25. Increment K
26. End For
27. For all k such that 0<k<(n - count) : Label2
28. If Var[k] < Min_Variance
29. Go to Label2
30. End If
31. New_columns[c] = k
32. Increment c
33. End For
34. Set P = [(n – count) – (c-1)]
35. Obtain the matrix A2[m][r] by removing
    columns in New_col from A1
36. Exit and output a Dataset Matrix A2[m][r],
    m=0: R, n= 0: P
```

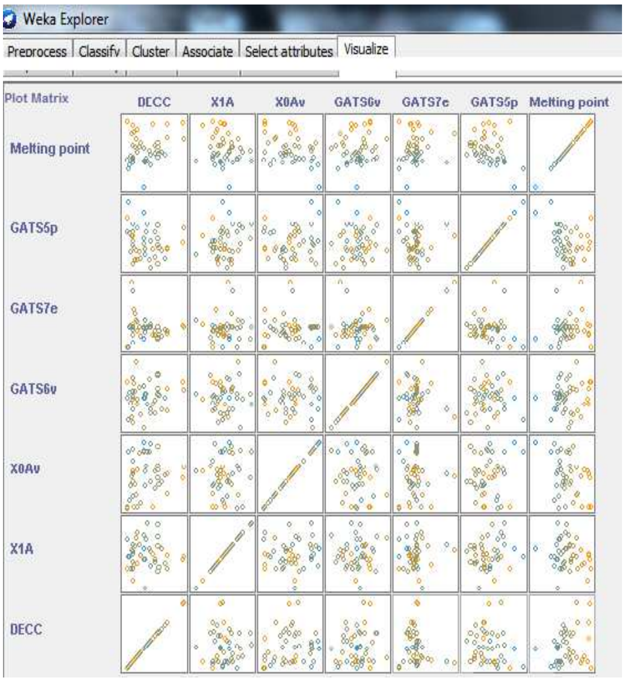

Fig 1. Correlation between selected descriptors.

4.3 Radial Basis Function Network Architecture

The prediction system is implemented using RBFN as shown in Figure 2 [15]. The input layer consists of total 6 neurons. Output layer has only one neuron whose output is the predicted melting point. Hidden layer neurons are chosen according to training data sample. The inputs for the RBFN are directly connected to each basis function. Then the output of the activation functions are weighted and summed. RBFN has the capability to transform nonlinear data input to linear output. The input to each hidden layer node is nonlinear in nature and it is treated with radial activation function. The final output is the weighted summation of these nonlinear inputs, thus transforming nonlinearity to linearity. RBFN have faster learning capacity, it is easy to implement, less complex in structure, and computationally more efficient.

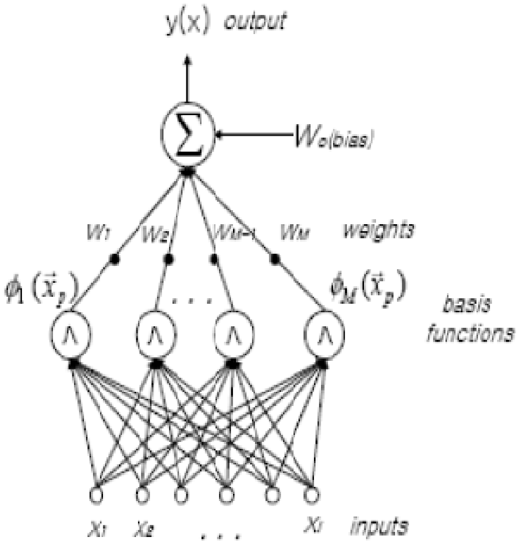

Fig. 2 RBNF Architecture

4.3 Radial Basis Activation Function

A radial basis function is a real valued function. Its value depends on the distance from the origin, so that
Ø(X) = Ø (||X||)

Every function Ø that satisfies this property is a radial function. In this study we implemented a normalized Gaussian radial basis function network.
$\varnothing(r) = e^{-(cr)^2}$ *where* $r = || x - x_i ||$

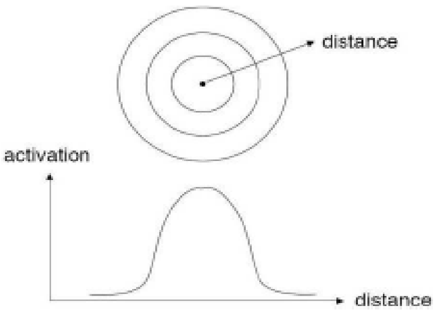

Fig. 3 Radial Basis Function

K-means clustering algorithm is used for designing of radial basis function network. Samples located far away from the cluster centers (mean) will fail to activate the Gaussian basis function. Data samples closest to a cluster's mean will achieve maximum activation.

4.4 RBFN Training

Supervised learning is employed to train RBNF model. The dataset was randomly divided into 2 groups – a training set consisting of 66 and a test set of 34 compounds. Training set was used for the model generation and test was used for the evaluation of the generated model.

5. RESULTS

The performance of the model was evaluated by Root Mean Square Error- RMSE which is defined as follows.

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{(P_i^{\exp} - P_i^{cal})^2}{N}}$$

Where $P_i^{\exp}$ and P_i^{cal} are the experimental and predicted/calculated values of melting point respectively. N denotes the number of data points. The RBFN predicted output showed an absolute error 12.5942 and RMSE of 15.8727. Certain outputs are given in Table 2. Actual and the corresponding predicted melting points are plotted in a graph and analyzed the RBNF performance. The processing of the data was carried out using WEKA Workbench 3.6.12.

Table 2. Actual Melting Point (AMP) VS Predicted Melting Point (PMP).

Si.No.	Medicine	AMP	PMP	Error
1	Bufexamac	153	144.65	- 8.35
2	Felbamate	151	157.87	6.867

3	Abecarnil	150	157.96	7.955
4	Haloperidol	148	157.79	9.785
5	Dapiprazole	158	153.31	-4.69
6	Erdosteine	156	142.99	-13.0
7	Glyburide	169	157.84	-11.1
8	Alpidem	140	143.31	3.307
9	Glisoxepid	189	157.93	-31.1
10	Carbutamide	144	144.67	0.674
11	Glyburide	169	156.34	-12.7
12	Famotidine	163	143.99	-19.0
13	Bezafibrate	186	157.50	-28.5
14	Actaminoselol	187	157.98	-29.0
15	Ahistan	144	145.89	1.89

6. CONCLUSION

In the present study, RBFN network trained with the supervised learning algorithm is used to predict the melting point of medicinal compounds. We designed and tested an RBFN classifier and obtained better results. Six vital influencing factors are input to an RBFN and melting point is obtained as output. This model could predict the melting point with a Mean Absolute Error 12.5942 and RMSE 15.8727. So we claim that an RBFN regression model is a promising candidate for QSAR studies and it also improves the computation time.

REFERENCES

[1]. Meylan, W. H.; Howard, P. H.; Boethling, R. S. (1996): *Environ. Toxicol.Chem. 15*, pp. 100.

[2]. Katritzky, A. R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Maran, U.; Karelson, M.(2001): *Cryst. Growth Des. 1*, pp. 261.

[3].Rahman, J.C.d.a.M.H..(1998): QSAR Approach To The Prediction Of Melting Points Of Substituted Anilines. 6th Int. Conf. on Mathematical Modeling, 11, pp. 843.

[4]. Bergstrom ,C.A.S; Norinder, U; Luthman K; Artursson, P.(2003): Molecular descriptors influencing melting point and their role in classification of solid drugs. J. Chem. Inf. Comp. Sci, pp. 1177-1185.

[5]. David, Sánchez; Antonio ,Moreno ; Web mining techniques for automatic discovery of medical knowledge; ,Dept. of Com. Sci. and Mathematics; Universita Rovirai Virgili (URV) Avda. Països Catalans, 26. 43007

[6].Yahya Hassanzadeh, Nazarabadi; Maje, modaresi; Bahram, Jafari; Sanaz Taheri ,Boshrooyeh.(2014): Predicting the melting point of organic compounds consist of Carbon,Hydrogen,Nitrogen and Oxygen using Multilayer Perceptron artificial neural networks.

[7]. Rahiman, Rafidha; Balakrishnan, Kannan; K B, Sherly(2012): Using Neural Network classifier Support Vector Machine Regression for the prediction of Melting Point of drug-like compounds.

[8]. Aziz ,Habibi-Yangjeh; Eslam, Pourbasheer; Mohammad ,Danandeh-Jenagharad(2007): Prediction of Melting Point for Drug-like Compounds Using Principal Component-Genetic Algorithm-Artificial Neural Network.

[9] .Shodhganga(2012): Estimation of Melting Point of small chemical and drug molecule.

[10]. M,Karthikeyan; Robert,C.Glen; Andreas, Bender(2007): General Melting Point Prediction Based on a Diverse Compound Data Set and Artificial Neural Networks.

[11]. Rahiman, Rafidha; Balakrishnan, Kannan; K B, Sherly (2013): Prediction of Melting Point of Drug-like compounds Using Neural Network Classifiers.

[12]. I V,Tetko; J,Gasteiger; R,Todeschini; A,Mauri; D,Livingston; P,Ertl; V A,Palyulin; E V,Radchenko; N S,Zefirov; A S,Makarenko; V Y, Thanchuk; V V,Prokopenko(2005): Virtual computational chemistry laboratory-design abd description; J,Comput.Aid.Mol.Des.

[13]. J, Zupan; J, Gasteiger(1999): Neural Networks in Chemistry and Drug Design.

[14]. I H, Witten; E, Frank; Data Mining(2005): Practical machine learning tools and techniques; Morgan Kaufmann; San Francisco.

[15]. Orr, Mark J(1996): Introduction to Radial Basis Function Networks.

[16]. MSDN.microsoft.com