Using Neural Network Classifier Support Vector Machine Regression for the prediction of Melting Point of Drug – like compounds

Rafidha Rahiman K.A,Kannan Balakrishnan
Department of Computer Applications
Cochin University of Science & Technology
Kochi, India
rafidharahiman494@gmail.com,
mullayilkannan@gmail.com

Sherly K.B
Department of Chemistry
Mar Athanasius College, Kothamangalam
Ernakulam, India
sherlykb@gmail.com

***Abstract*— In our study we use a kernel based classification technique, Support Vector Machine Regression for predicting the Melting Point of Drug – like compounds in terms of Topological Descriptors, Topological Charge Indices, Connectivity Indices and 2D Auto Correlations. The Machine Learning model was designed, trained and tested using a dataset of 100 compounds and it was found that an SVMReg model with RBF Kernel could predict the Melting Point with a mean absolute error 15.5854 and Root Mean Squared Error 19.7576**

***Keywords*— Data Mining; Machine Learning; Support Vector Machine; QSA; Melting Point.**

I. Introduction

Drug design is the process of identifying new medicines. Drug is an organic molecule which inhibits or activates the function of a bimolecule and gives therapeutic benefits to the patients. Different compounds are used to make a particular drug. So it is important to know about the physical properties of such compounds like solubility, viscosity, Melting Point etc. Melting Point is a fundamental physical property used for screening and purity analysis. Since Melting Point affects Solubility, Viscosity and Toxicity of a compound, it is used as a descriptor for predicting these properties [1]. So the compound has to be synthesized and melting point has to be determined experimentally before predicting such properties. Here comes the importance of a computational model for predicting the Melting Point which is very much economical. Melting Point is calculated in laboratories very fast and straight forward but sometimes compounds used for this process may be hazardous, toxic, and expensive and may lead to the wastage of time and chemicals. So a computational model for predicting the Melting Point is desirable.

Data Mining is the process of discovering patterns and establishing relationships automatically or semi automatically to predict the future behaviour using machine learning techniques [2]. Machine Learning is the ability of a machine to automatically learn to recognize patterns by using Artificial Intelligence and make intelligent decisions based on the pattern. Many methods like Classification, Association and Clustering have been used for Machine Learning and Artificial Neural Network (ANN), Radial Basis Function Neural Network (RBFNN), K Nearest Neighbor (KNN), Clustering and Random Forest have been used in the past for these purpose. See for example [2][3][4].Also used SVM for predicting the physical properties Solubility, Melting Point and Log P[5] In this paper we use Support Vector Machine (SVM), a Supervised Learning method for Classification via Regression for predicting the Melting Point.

Molecular descriptors are molecular properties used to characterize the molecule. They are the result of logical and mathematical procedures which transforms chemical information in symbolic representation into the result of standard experiment. Molecular descriptors [6] helps to predict the activity and properties of molecules in complex experiments and play an important role in scientific growth [7].The molecular descriptors are distinguished by their physico chemical meaning or specific mathematical tools used for calculating molecular descriptors. QSAR/QSPR is based on the assumption that compounds from same chemical domain behave similarly. The general idea behind this is to get the physical property of the compound using a predictive model with the use of a database of values obtained from the previous laboratory experiments.

II. Data

A. Methodology

In our study, we used the Melting point data set of 100 drug - like compounds compiled by Bergestrom [1] [8]. The data set contains compounds with their melting points which was experimentally determined, and a SMILES (Simplified Molecular Input Line Entry System) description (Molecular structure specification format) of the compounds, which is used for calculating descriptors. The methodology used was to calculate the descriptors and to select the appropriate descriptors. E-Dragon[8] ,an electronic remote version of the Dragon software was used for calculating molecular descriptors.

978-1-4244-7926-9/11/$26.00 ©2011 IEEE

B. Data Analysis

Many factors will affect the success of Data Mining algorithms. The quality of data is a major concern. If dataset contains irrelevant and redundant information or data is noisy or unreliable then it is a more difficult process to select the attributes. So the attributes (Descriptors) that have constant value 0 is eliminated in a pre reduction step to avoid redundancy [2]. Attribute selection is a process of identifying and removing irrelevant and redundant information from the dataset so that training algorithms operate faster and effectively. If user has background knowledge about data then he can select his own dataset and this is better than using attribute selection methods [9].The descriptors for hydrophilicity, polarity, partial atomic charges and molecular rigidity are positively correlated with melting point[1].Finally 7 descriptors are selected and given in the table1. Figure1 shows the correlation of these descriptors with Melting Point. After that the compound which has melting point in between 130^0c and190^0 c [1] is chosen.

III. MODEL SELECTION

After selecting the descriptors, we split the dataset into a training set and a test set, setting aside 66% of the data into the training set and the remaining into the test set .The split was used to evaluate the performance of the classifier. Then an SVMReg model is designed for prediction.

TABLE I. THE DESCRIPTORS USED

pol	Polarity Number
Unip	Unipolarity
Har	Harary H index
PHI	Kier flexibility index
LPRS	Log of product row sums
Hydp	Hyper distance path index
nH	Number of Hydrogen atoms

SVM is a group of supervised learning methods used for classification and regression. Classification is a Data Mining technique used to predict the target value after seeing a number of training samples. The actual outcome is given as the class for each instance in the input data and the class is nominal.

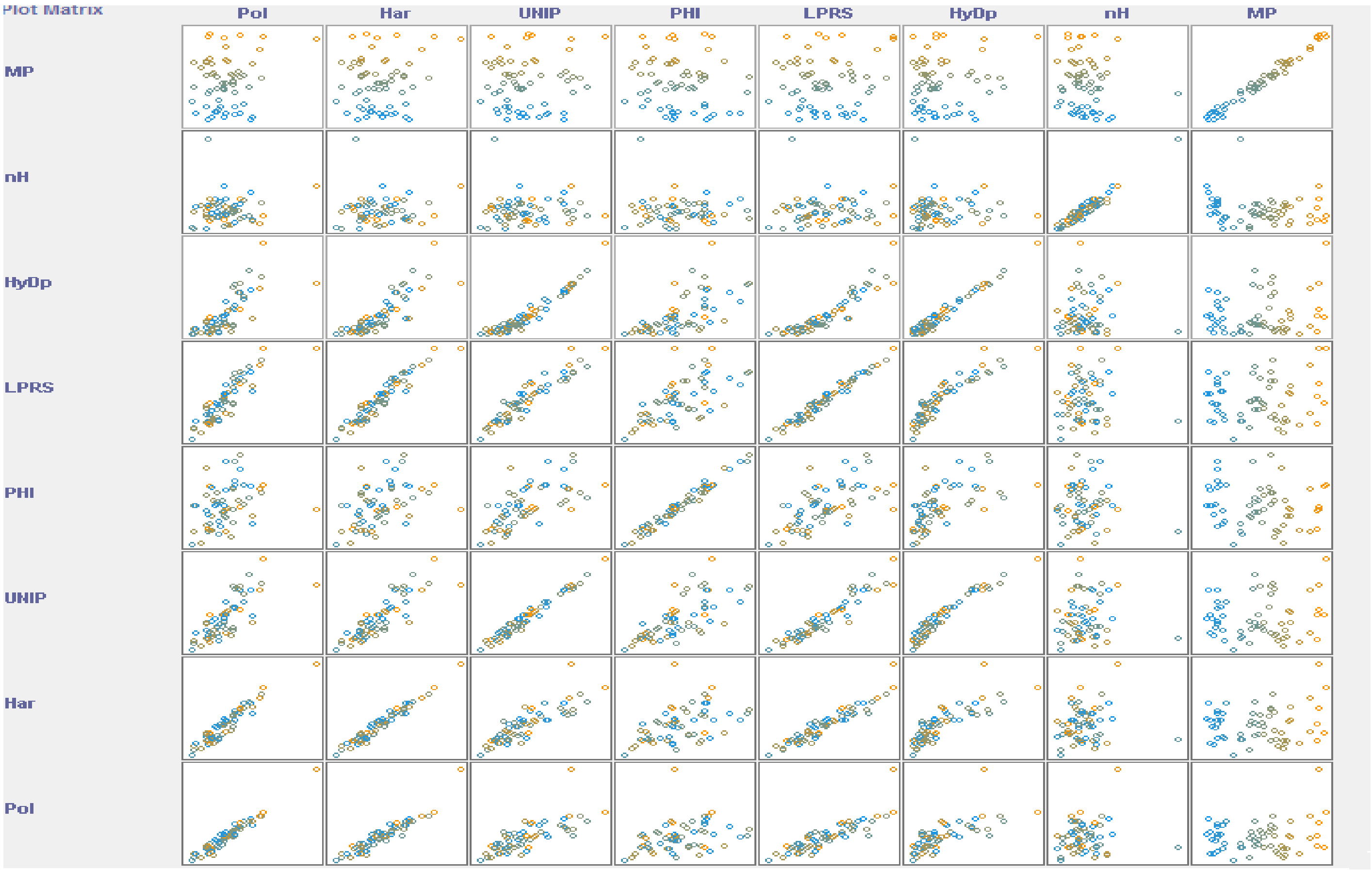

Figure 1. Correlation between selected descriptors

SVM tries to obtain a hyperplane with maximum Euclidian distance to the nearest neighbour point to optimally separate different classes. But in case of real data sometimes the class or all the attribute may be numeric and we apply classification via regression also known as Linear classification or logistic regression to predict the future behavior. Linear regression finds a function that approximate the training points by minimizing prediction error. When minimizing error, the risk of overfitting is reduced as algorithm simultaneously tries to maximize the flatness of regression function. Also it minimizes the predictions absolute error. SVM maps lower dimensional input vectors into a high dimensional feature space by using a kernel function and then construct an optimal separating hyperplane[2]. Basic kernels available with SVM are Linear, polynomial, Radial basis function and Sigmoid and among all these RBF is the first reasonable choice [10] and it will minimize the root mean square error. We then used Weka[2] (Waikato Environment for Knowledge Analysis), a powerful open source Java based Data Mining tool to build the model and analyze the results. Weka can be run on any computer that has a Java run time environment. It is a framework with graphical user interface and brings together almost all machine learning algorithms and tools. Here we designed an SVMreg model in Weka with a Gaussian Radial Basis Function Kernel. An SVM model with RBF kernel and c (complexity parameter) value 1 is found to be most optimal.

The predicted output showed an absolute error 15.584 and RMSE 19.7576 and it was better than KNN, K *, RBFNN (Radial Basis Function Neural Network) and BPNN (Back Propagation Neural Network).

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} (Xi - Yi)^2 / n}$$

And

$$\text{Absolute Error} = \sum_{i=1}^{n} |Xi - Yi| \;/n$$

Where X_i is the actual value, Y_i is the predicted value and n is the number of molecules.

Figure 2 shows the SMOreg(Sequential minimal optimization algorithm for support vector regression) knowledge flow model designed in Weka, Table 2 shows the results of SVM in comparison with other models[11][12] and Table 3 shows the actual melting point, melting point predicted by SVMreg model and the relative error.

TABLE II. COMPARISON OF SVM WITH OTHER MODELS

SVM		**KNN**		**K ***		**RBFNN**		**BPNN**	
M A E	**RMSE**	**M A E**	**RMSE**	**M A E**	**RMSE**	**M A E**	**RMSE**	**M A E**	**RMSE**
15.584	19.7576	27.52	39.69	30.409	43.997	40.4627	49.7965	48.908	63.5923

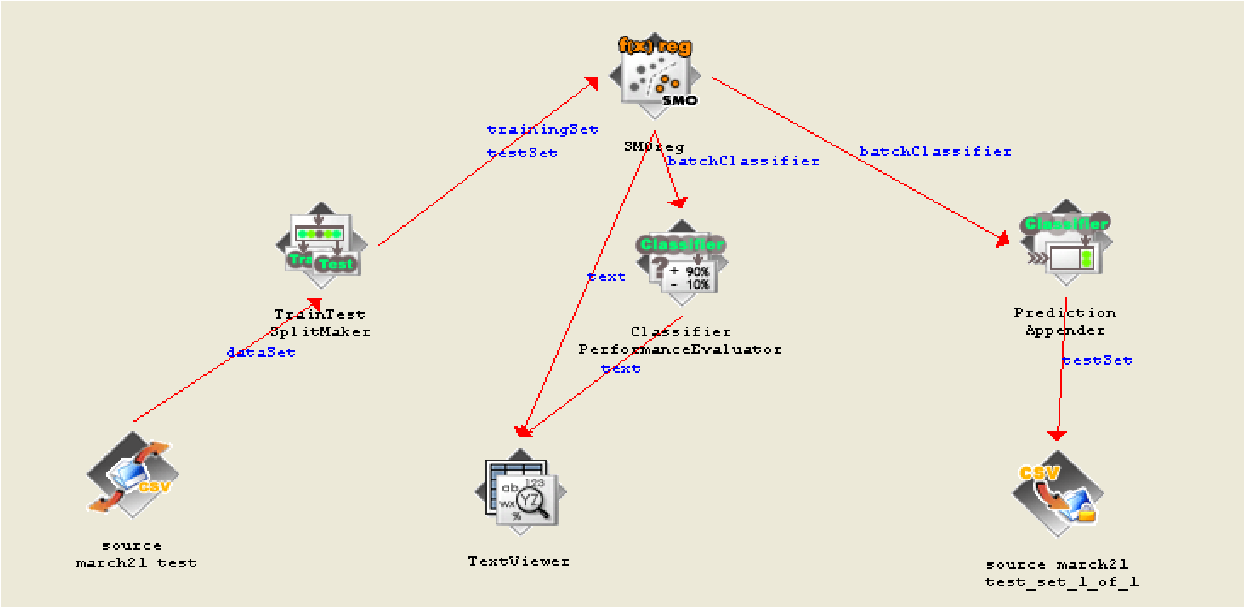

Figure 2 SVM Knowledge flow model designed in weka

TABLE III. ACTUAL MELTING POINT MELTING POINT PREDICTED BY THE MODEL AND RELATIVE ERROR

Name of compound	Actual Melting Point	class_ predicted_by: SMOreg'	Relative Error
Antazoline	120	139.892053	-17
Difenpiramide	122	142.80302	-17
Erdosteine	156	143.370372	8
Atenolol	146	140.811644	4
Abecarnil	150	143.652444	4
Chlophedianol	120	138.630235	-16
Bamipine	114	141.620284	-24
Flumetramide	115.5	139.687828	-21
Florfenicol	153	143.638384	6
Felodipine	145	140.232424	3
Azacyclonol	160	142.853851	11
Aminorex	136	143.485826	-6
Benzarone	124.3	140.508029	-13
Acemetacin	150	138.779658	7
Carbutamide	144	141.227186	2
Benzydamine	160	144.722679	10
AcetylsalicylicAcid	142.4	143.766928	-1
Benperidol	170	138.886979	18
Acifran	176	143.070555	19
Glisoxepid	189	139.294836	26
EthinylEstradiol	141	139.627829	1

IV. CONCLUSION

In this paper in order to predict the Melting Point of Drug – like compounds, we designed and tested an SVMreg classifier and obtained better results when compared with KNN (K Nearest Neighbour) model [10].SVMreg model with RBF Kernel could predict the Melting Point with a mean absolute error *15.5854* and Root Mean Squared Error *19.7576*.So we claim that an SVM Regression model with RBF Kernel is a promising candidate for QSAR studies and it improves the computation time. The study emphasized the need for more suitable attributes (descriptors) to design the model to predict the physical properties.

REFERENCES

[1] C.A.S. Bergstrom, U. Norinder, K. Luthman and P. Artursson, "Molecular descriptors influencing melting point and their role in classification of solid drugs", J. Chem. Inf. Comput. Sci, 43(4), pp.1177-1185 (2003):

[2] I. H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[3] Du QS, R.B. Huang and K.C. Chou, "Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design", Guangxi University, Key Laboratory of Subtropical Bioresource Conservation and Utilization of Guangxi, Nanning, Guangxi, 530004, China.

[4] H. Liu, X. Yao and P. Gramatica, "The applications of machine algorithms in the modeling of Estrogen like chemicals", School of Pharmacy, Lanzhou University, Lanzhou 730000, China.

[5] Laura D.Hughes, David S Palmer,Florian Nigsch and John B.O Mitchell, "Why are some properties difficult to predict than others ? A study of QSPR models of solubility, melting point and log P",Journal of chemical information and modelling,.

[6] Pang Ning Tan, Michael Steinbach and Vipin Kumar, "Introduction To Data Mining".

[7] Roberto Todeschini and Viviana Consonni, "Handbook of molecular descriptors, Wiley – VCH, 2000.

[8] I. V. Tetko, J Gasteiger, R. Todeschini, A. Mauri, D. Livingstone,P. Ertl, V.A. Palyulin, E.V. Radchenko, N.S Zefirov, A.S.Makarenko,V.Y. Tanchuk and V.V.Prokopenko,"Virtual computational chemistry laboratory -design and description", J. Comput. Aid. Mol. Des., 2005, 19, 453-63

[9] Mark A. Hall and Geoffrey Holmes, "Benchmarking attribute selection technique for discrete class data mining".

[10] Chih Wei Hsu, Chih Chung Chang and Chih Jen Cin, "A practical guide to support vector classification".

[11] Kannan Balakrishnan, Sherly K.B and Rafidha Rahiman K.A,"Prediction of melting point using neural network classifiers KNN and K*", Proceedings of national conference on Computational Chemistry 2009 , M.A College Kothamangalam.

[12] Rafidha Rahiman K.A, Kannan Balakrishnan and Sherly K.B,"Using neural network classifiers for predicting the Melting Point of Drug – Like compounds", Proceedings of national conference on Soft Computing 2010 , Marian College Kuttikanam.