QSPR OUTLIER DETECTION METHODS

*B. Firdaus Begam**

Abstract

The high performance of QSPR may get affected when the dataset contains outlier data which leads to misleading output. Occurrences of outlier in dataset are caused by various reasons like human error, change in the method of collecting data and data processing methods. When generating QSPR model for high dimensional data, there may be possibility of getting fault result, if outliers are not detected and handled before analysis. There occurs the situation where outlier detection and removal of outlier to be done as pre-processing step to develop a better QSPR prediction model.

Keywords : Outlier, univariate, multivariate.

I. INTRODUCTION

The fast development of computing technologies in today's world has changed the method of acquiring, storing, maintaining and manipulating data. Information technologies used to solve chemical problem has given rise to new field called "Cheminformatics" or "Chemical Informatics". According to FK.Brown cheminformatics is termed as "the use of information technology and management has become a critical part of the drug discovery process"[1]. M.Hann & R. Green termed Cheminformatics as a new name for an old problem [2].

The chemical information of molecules was represented based on graph theory and it is also known as chemical graph theory. The nodes represent the atoms and the edges represent the connectivity (bond) between the atoms [3]. The molecular structure information is represented by various chemical mol file. Based on chemical graphical representation various molecular descriptors are calculated. The processes of calculating or predicating property of a molecule based on structural information are called as chemmetrics.

The term chemometrics is analogy with biometrics, econometrics, etc. is heavily dependent on the use of different kinds of mathematical models to predict property of molecule based on its structural information [4]. Chemmetrics or chemometrics is used for quantitative analyse of the chemical data by using mathematical and statistical methods [5]. Chemometrics is the branch of chemistry concerned with the analysis of chemical data (extracting information from data) and ensuring that experimental data contain maximum information (the design of experiments) (Figure 15).

According to Massart, chemical problems that solved using mathematics and statistical logic are termed as chemometrics. It can be use to defined experimental procedure how to design and select optimal methods, provides more relevant and efficient information to analyze chemical data which helps in obtaining maximum knowledge/information about chemical systems.
Chemometrics overlaps with cheminformatics along with machine learning algorithm to analyze data. Some areas of applications are analytical chemistry, chemical engineering and other related fields [6].

II. QUANTITATIVE STRUCTURE PROPERTY RELATIONSHIP (QSPR)

Quantitative Structure Property Relationship/ Quantitative Structure Activity Relationship (QSAR/ QSPR)

Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
*Corresponding Author

prediction model was first introduced by Hansch to understand relationship that exist between chemical/ molecular structure of a chemical compound molecules to its properties like biological, physical and chemical property [7].

QSPR are a mathematical model which involves various steps, acquiring chemical data (Experimental /Observation), based on property to exhibit dataset are selected for chemical (mathematical) analysis and various statistical analysis and correlation are calculated (Combinatorial techniques) (Figure 1). Based on significance and performance the methods which provides best results are used to develop prediction model.

A computational method describing Structure Property Relationship (SPR) quantitatively is called quantitative structure-Property relationship (QSPR). QSPR/QSAR research has gained importance over recent decades as the predication ability over various properties (physical, biological, chemical, ADME, toxicological and other related properties) has shown its influence in various fields of application like pharmaceutical, agriculture, dying, textile, drug discovery [8,9].QSPR/QSAR model expresses the characteristics of molecule through various molecular descriptors like one-dimensional, two dimensional and three-dimensional descriptors".

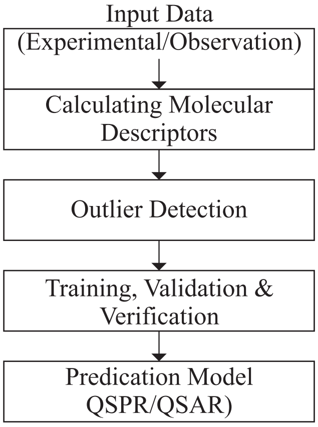

Figure 1: QSPR Approach

III. OUTLIER DETECTION

An Observation present in a pool of data showing inconsistency are outliers present in dataset. According to Barnett and Lewis, outliers in dataset are to be "an observation or subset of observations which appears to be inconsistent with the reminder. When an observation performs or hold a different pattern compared to other remaining data present in dataset are considered as Outliers [10].

The presence of outliers may occur under various circumstances like, due to presence of variability in dataset, as an outcome of an error or the data acquired from other model. The presence of inconsistent data fails to follow a general pattern as a result it has a greater impact on variance and correlation among the variables present in the dataset. Thus, it is critical and important to handle such type of data by detecting and treating contaminated observations, as it leads or contributes to degrading the performance of dataset. When an observation in a dataset deviates from other observations, it results it to be suspicious data calculated using different methods is known as outlier.[11]

Common reasons of outliers:

1. Human error caused due to error made during data entry.
2. Data collection error caused by measurement or experimental errors
3. Data manipulation error caused during data processing errors.

Univariate outlier detection methods:

When the relationship between chemical descriptor and observation in a dataset deviates from other dataset is known as outlier. Some of univariate outlier detection methods are as follows,

- Standard Deviation
- ZScore
- Median absolute deviation

Multivariate outlier detection:

Three different types of outliers can be defined under multivariate statistics are,

1. When chemical or molecular descriptors (X) and individual molecules (Y) is not same as other data in training set.
2. When a particular molecular descriptor (X) deviated from the pattern of other descriptors.
3. When a particular molecule(s) (Y) generates invalid response.

Multivarate methods:

- Using Mahalanobis Distance
- DBScan clustering method
- Isolation Forest
- Regression methods
- Minmax and Maxmin algorithms
- Fuzzy based MinMax and Maxmin algorithms

REFERENCES

[1] Brown FK. Chemoinformatics: what is it and how does it impact drug discovery? Ann Rep Med Chem; 1998, 33:375 384.

[2] Hann, M., Green, R. Chemoinformaticss A new name for an old problem. Curr. Opin. Chem.Biol.; 1999, 3, 379-383.

[3] A. T. Balaban and D. H. Rouvray, Chemical applications of graph theory. In Applications o f Graph Theory (eds. R. J. Wilson and L. W. Beineke). Academic Press, New York (1979) 177-221.

[4] SvanteWold, Chemometrics; what do we mean with it, and what do we want from it?, Volume 30, Issue 1, November 1995, Pages 109-115.

[5] SvanteWold, MichaelSjöström, LennartEriksson, PLS-regression: a basic tool of chemometrics, Volume 58, Issue 2, 28 October 2001, Pages 109-130.

[6] Lucien Birgé. Pascal Massart. "Minimum contrast estimators on sieves: exponential bounds and rates of convergence." Bernoulli, 4(3) 329-375 sept 1998.

[7] Hansch C, Leo A, Mekapati SB, Kurup A. QSAR and ADME. Bioorganic & Medicinal Chemistry. 2004; 12(12):3391–400.

[8] Faulon JL, Bender A. Handbook of cheminformatics algorithm; CRC press: New York; 2010.

[9] Leach AR, Gillet VJ. An introduction to chemoinformatics; Springer Science Business Media Inc: India; 2007.

[10]Barnett, V and Lewis, T. (1995), Outliers in Statistical Data, Chichester: John Wiley and Sons. pp.7, 269.

[11] Williams, R. Baxter, H. He, S. Hawkins and L. Gu, “A Comparative Study for RNN for Outlier Detection in Data Mining”. In Proceedings of the 2nd IEEE International Conference on Data Mining, page 709, Maebashi City, Japan, December 2002.