International Journal of Machine Intelligence
ISSN: 0975-2927 & E-ISSN: 0975-9166, Volume 7, Issue 1, 2016, pp.-469-473.
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BPJ0000231

Research Article

DECISION TREE LEARNING AND REGRESSION MODELS TO PREDICT ENDOCRINE DISRUPTOR CHEMICALS - A BIG DATA ANALYTICS APPROACH WITH HADOOP AND APACHE SPARK

PAULOSE RENJITH[1]*, JEGATHEESAN K.[2] AND GOPAL SAMY B.[3]

[1]Cognizant Technology Solutions, Athulya, Infopark SEZ, Kakkanad, Kochi 682030, Kerala, India
[2]Center for Research and PG Studies in Botany and Department of Biotechnology, Thiagarajar College (Autonomous), Madurai - 625 009, Tamil Nadu, India
[3]Department of Biotechnology, Liatris Biosciences LLP, Kakkanad, Cochin, Kerala
*Corresponding Author: Email-renpau@gmail.com

Received: April 04, 2016; Revised: May 04, 2016; Accepted: May 05, 2016; Published: May 07, 2016

Abstract- Predictive toxicology calls for innovative and flexible approaches to mine and analyse the mounting quantity and complexity of data used in it. Classification and regression based machine learning algorithms are used in this study in order to computationally predict chemical's affinity towards endocrine hormones. As a result of the modelling complexity and existing big sized toxicity datasets generated by various irrelevant descriptors, missing values, noisy data and skewed distribution, we are motivated to use machine learning and big data analytics in toxicity prediction. This paper reports results of a qualitative and quantitative toxicity prediction of endocrine disrupting chemicals. Datasets of Estrogen Receptor (ER) and Androgen Receptor (AR) disrupting chemicals along with their Binding Affinity values were used for building the predictive models. Fragment counts of dataset chemicals were generated using Kier Hall Smarts Descriptor that exploit electro-topological state (e-state) indices. Chemical data after fingerprint calculations were loaded into Hadoop Distributed File System (HDFS) for parallel processing. Decision tree learning classifier algorithm was applied using Apache Spark big data processing framework to qualitatively predict endocrine disruptor and non-disruptor chemicals. ER and AR predictive models over training datasets demonstrated 89.5% and 90.03% accuracy in toxicity prediction whereas corresponding models on their test datasets showed 81.25% and 73.33% prediction accuracies respectively. Linear regression algorithm built using R statistical software was used to quantitatively predict the log Relative Binding Affinity (logRBA) of chemicals towards Androgen and Estrogen Receptors. This study details the power of Decision Tree Learning algorithm in chemical toxicity prediction on a Hadoop parallel computing environment that can be leveraged to explore advanced machine learning technologies for getting high accuracy in chemical toxicity prediction.

Keywords- Endocrine Disruptor, Chemical Toxicity Prediction, Decision Trees Learning Algorithm, Machine Learning, QSAR, Big Data, Apache Spark, Hadoop, Linear Regression Algorithm, Qualitative Prediction, Quantitative Prediction.

Citation: Paulose Renjith, et al., (2016) Decision Tree Learning and Regression Models to Predict Endocrine Disruptor Chemicals - A Big Data Analytics Approach with Hadoop and Apache Spark. International Journal of Machine Intelligence, ISSN: 0975-2927 & E-ISSN: 0975-9166, Volume 7, Issue 1, pp.-469-473.

Academic Editor / Reviewer: Kadchha Rajeshbhai, Ravisankar H., Jayakrishna Yogesh, Raj Deep Singh Shaktawat

Introduction

Development of cosmetic products by means of animal testing might engage trying either a complete product or its individual ingredients on animals like mice, rabbits and rats. Cosmetic testing on animals has been banned in many countries owing to the strong public backlash against it. India became the second country in Asia by declaring a ban on testing cosmetics on animals in the country in early 2014 [6]. Soon after in November 2014 India prohibited importing cosmetics tested on animals [5].

Yet, they are still essential to show that the products and/or their ingredients are harmless and safe to public health. A cutback in the quantity of animals employed and the testing perfection to decrease anguish are significant objectives for the concerned industries [4]. Executions of improved test methods that stay away from the involvement of live animals are referred to as alternatives to animal testing. *In-vitro* cell culture techniques and in silico computer simulation are the two major alternatives to *in-vivo* animal testing.

An Artificial Intelligence element called Machine Learning constantly monitors a sequence of activities executed in a particular time and utilizes this knowledge to work out the methods to carry out analogous processes better in a new situation. Machine Learning was defined as the speciality that provided computers the capability to gain knowledge devoid of explicit programming [9].

While the Machine Learning field is still budding these days, it has made its availability into daily customer practice through applications similar to Google Maps[1] that provide precise geographical data even up to street view and Netflix [2] that suggest the client practice through movie viewing habit prototypes. Applications used for advertisement placement, credit scoring, drug design, facial recognition (iPhoto), fraud detection, gesture recognition (Kinect), spam filters, speech and natural language processing (Siri), stock trading and web search are other examples.

The growth of Machine Learning applications is accelerated by Big Data technology that has the ability of processing huge volumes of data. A compilation of large datasets that are unable to process by conventional computing techniques is called Big Data. It includes extensible variety, elevated velocity, and enormous volume of data that has turned out to be a complete subject involving a variety of frameworks, techniques and tools.

There are a variety of technologies to handle big data in the market from different traders like Microsoft, IBM, and Amazon and these technologies are of two classes namely Operational Big Data and Analytical Big Data. Systems like

MongoDB that offer operational capabilities for synchronized interactive workloads comes under Operational Big data and their data is mainly confined and stored. Analytical Big data have Massively Parallel Processing (MPP) database systems and Map Reduce that supply complex analysis and analytical retrospective capabilities that might handle almost all of the data.

Map Reduce presents a novel data analysing method explicitly harmonizing to the abilities offered by SQL. Applications processing large amount of data in parallel with large clusters of service hardware in a consistent, fault-tolerant approach is easily written by a software framework called Apache Hadoop Map Reduce.

Apache Mahout is a machine learning collection for Hadoop that contains a scalable Machine Learning algorithm collection implemented in rapid cycles with the ground-breaking Map Reduce technology [10]. These algorithms have an outstanding capability to craft predictions and reveal hidden relationships among datasets.

Trendy means for the machine learning responsibilities of regression and classification are Decision trees and their ensembles. Decision trees are broadly employed as they are easy to handle categorical features, pull out to the multiclass classification setting, interpret, do not necessitate feature scaling, and are capable of capturing feature interactions and nonlinearities. Decision trees in data mining are the combination of computational and mathematical techniques to support the classification, explanation and simplification of a given dataset. Constructing decision tree from class labeled training tuplesis called as Decision tree learning. A decision tree is a flowchart resembling structure with every internal node representing an investigation on an attribute, each branch indicating a test result and each terminal node holding a class label.The pinnacle performers for regression and classification tasks are Tree ensemble algorithms like random forests and boosting. Spark MLlib upholds decision trees for regression and for binary and multiclass classification by means of both categorical and continuous features.The objective is to generate a model that forecast the target variable value anchored in numerous input variables.

Chemicals that get in the way with the endocrine systems resulting in unfavourable effects are called Endocrine disruptors. The estrogens and androgen receptors are such chemicals achieving this by binding to receptors. These chemicals are also referred as endocrine active compounds, environmental hormones and endocrine modulators. A wide and varied range of substances including DDT, diethylstilbestrol or the synthetic estrogen DES, polychlorinated biphenyls (PCBs), dioxin and dioxin-like compounds and some other pesticides are thought to cause endocrine disruption.

Decision tree learning is a frequently used predictive modelling method in machine learning and data mining [8]. A decision tree was used as a prophetic model in decision tree learning to plot annotations on an item to conclusions concerning the item's objective worth. Classification trees are the tree models with target variable having a fixed set of values and decision trees with target variable having continuous values are known as regression trees. A decision tree in decision study was employed to represent decisions openly. A decision tree in data mining depicts only data and not decisions but the ensuing classification tree was a key for making decision.

In the current study, the role and power of Decision tree learning algorithm over prediction of Endocrine disruption nature of chemicals is explored.

Materials and Methods

Data and Software Availability

Androgen Receptor and Estrogen Receptor binding chemical datasets were obtained in SDF format from the FDA Endocrine Disruptor Knowledge Base website. Chemical structures along with their logRBA values against AR and ER were used for statistical predictive model building and evaluation. The open source pipeline generation platform KNIME v.2.10.0 was used for the data preponderance tasks. Hadoop 2.7.1 and Apache Spark 1.4.0 technology frameworks were used for statistical model building on a Big Data platform. R statistical software was used for regression model building and quantitative prediction.

Data Preparation

All molecular structures were standardized and compounds with ambiguous activity values as well as duplicates were removed in the first step. A total of 171AR and 232 ER binding chemical data were taken for consideration. A Spreadsheet was prepared with molecule IDs and corresponding log Relative Binding Affinity (logRBA) values. Chemicals were qualitatively labeled as Disruptor or Non-Disruptor based on the following logRBA ranges. Chemicals with logRBA value ranges between -4.5 to 2.6 were classified as Disruptors and with value of -10,000 were classified as Non-Disruptors. With these conversions, a new spreadsheet was developed.

Training and Test Datasets

Datasets were randomly separated into training and test sets for statistical model building and evaluation purpose. ER dataset was divided into 200 molecules of training set and 32 molecules of test set. Similarly, AR dataset comprised of 160 training molecules and 34 test molecules. Separate spreadsheets were created for AR and ER training and test datasets.

Molecular Fingerprint Calculations

Kier Hall Smarts Descriptor that uses electro-topological state (e-state) indices was generated for each molecules using RCDK package in R statistical software. Spreadsheets were amended with structural descriptors and separately maintained for AR and ER training and test datasets.

Hadoop and Apache Spark

Chemical data after fingerprint calculations were loaded into HDFS for Parallel Processing. Decision tree learning algorithm was applied over AR and ER training datasets using Apache Spark MLlib and predictive models were generated. These predictive models were applied over the test datasets and endocrine disrupting and non-disrupting chemicals were predicted and qualitatively classified. The prediction accuracies were calculated separately for each datasets.

Linear Regression for Quantitative Prediction

Linear regression algorithm was applied over training data using R statistical software and predictive models were built separately for AR and ER training datasets. These models built were applied over test datasets and log Relative Binding Affinity (logRBA) values towards AR and ER were predicted quantitatively. The prediction accuracies were calculated for each datasets.

Results and Discussion

Qualitative Prediction Using Decision Tree Algorithm

The training and test datasets utilized for this study were already utilized in benchmarking the classification algorithms in our previous work [7]. The decision tree models generated for both estrogen receptor (ER) and androgen receptor (AR) disruptor datasets are shown in [Table-1]. The results of decision tree algorithm applied on both estrogen receptor (ER) and androgen receptor (AR) training datasets are given in [Fig-1]. Decision tree learning model on AR training dataset results 73.05% true positives, 17.02% true negatives, 9.22% false positives, and 0.71% false negatives. Whereas for ER training set, 53% true positive, 36.5% true negatives, 8% false positives, and 2.5% false negatives were resulted. [Fig-2] depicts the results of decision tree algorithm applied to both ER and AR test datasets. AR test dataset resulted 60% true positives, 13.33% true negatives, 23.33% false positives, and 3.33% false negatives. Whereas for ER test dataset, 43.75% true positives, 37.5% true negatives, 0% false positives, and 18.75% false negatives resulted. These decision tree learning models built using Apache spark demonstrate efficient accuracy in chemical toxicity prediction similar to the earlier works in Spark [12].

Quantitative Prediction Using Linear Regression

R statistical software is used to build the linear regression models with ER and AR datasets. Datasets in HDFS were accessed and analyzed using R-Hadoop package in R statistical environment. Linear regression model with dataset gave Residual standard error: 1.3, Multiple R-squared: 0.5749, F-statistic: 6.911, and p-value: $1.156e^{-10}$. Whereas AR model gave Residual standard error: 0.9798, Multiple R-squared: 0.5353, F-statistic: 3.548, and p-value: $1.048e^{-05}$. The graph of actual and predicted logRBA value for ER training set and AR training set using linear regression is given in the [Fig-3] and [Fig-4] respectively. The actual and

Table-1 *Decision tree model generated for ER and AR datasets*

ER	AR
Decision Tree Model classifier of depth 5 with 37 nodes	Decision Tree Model classifier of depth 5 with 29 nodes
If (feature 33 <= 0.0)	If (feature 17 <= 0.0)
If (feature 11 <= 10.0)	If (feature 15 <= 1.0)
If (feature 18 <= 3.0)	If (feature 11 <= 0.0)
If (feature 18 <= 0.0)	If (feature 12 <= 2.0)
If (feature 6 <= 3.0)	Predict: 0.0
Predict: 0.0	Else (feature 12 > 2.0)
Else (feature 6 > 3.0)	If (feature 6 <= 2.0)
Predict: 0.0	Predict: 1.0
Else (feature 18 > 0.0)	Else (feature 6 > 2.0)
If (feature 11 <= 3.0)	Predict: 0.0
Predict: 0.0	Else (feature 11 > 0.0)
Else (feature 11 > 3.0)	If (feature 16 <= 1.0)
Predict: 0.0	Predict: 0.0
Else (feature 18 > 3.0)	Else (feature 16 > 1.0)
If (feature 15 <= 0.0)	If (feature 8 <= 1.0)
Predict: 0.0	Predict: 1.0
Else (feature 15 > 0.0)	Else (feature 8 > 1.0)
Predict: 1.0	Predict: 1.0
Else (feature 11 > 10.0)	Else (feature 15 > 1.0)
If (feature 15 <= 0.0)	If (feature 33 <= 1.0)
Predict: 0.0	If (feature 12 <= 4.0)
Else (feature 15 > 0.0)	Predict: 1.0
Predict: 1.0	Else (feature 12 > 4.0)
Else (feature 33 > 0.0)	If (feature 8 <= 7.0)
If (feature 16 <= 1.0)	Predict: 1.0
If (feature 8 <= 8.0)	Else (feature 8 > 7.0)
Predict: 0.0	Predict: 0.0
Else (feature 8 > 8.0)	Else (feature 33 > 1.0)
If (feature 6 <= 2.0)	If (feature 16 <= 4.0)
If (feature 12 <= 5.0)	If (feature 34 <= 2.0)
Predict: 0.0	Predict: 1.0
Else (feature 12 > 5.0)	Else (feature 34 > 2.0)
Predict: 1.0	Predict: 0.0
Else (feature 6 > 2.0)	Else (feature 16 > 4.0)
Predict: 0.0	Predict: 0.0
Else (feature 16 > 1.0)	Else (feature 17 > 0.0)
If (feature 33 <= 4.0)	If (feature 15 <= 0.0)
If (feature 8 <= 1.0)	Predict: 0.0
If (feature 11 <= 3.0)	Else (feature 15 > 0.0)
Predict: 0.0	If (feature 6 <= 1.0)
Else (feature 11 > 3.0)	Predict: 1.0
Predict: 1.0	Else (feature 6 > 1.0)
Else (feature 8 > 1.0)	Predict: 0.0
If (feature 20 <= 0.0)	
Predict: 1.0	
Else (feature 20 > 0.0)	
Predict: 0.0	
Else (feature 33 > 4.0)	
If (feature 17 <= 0.0)	
Predict: 0.0	
Else (feature 17 > 0.0)	
If (feature 11 <= 5.0)	
Predict: 1.0	
Else (feature 11 > 5.0)	
Predict: 0.0	

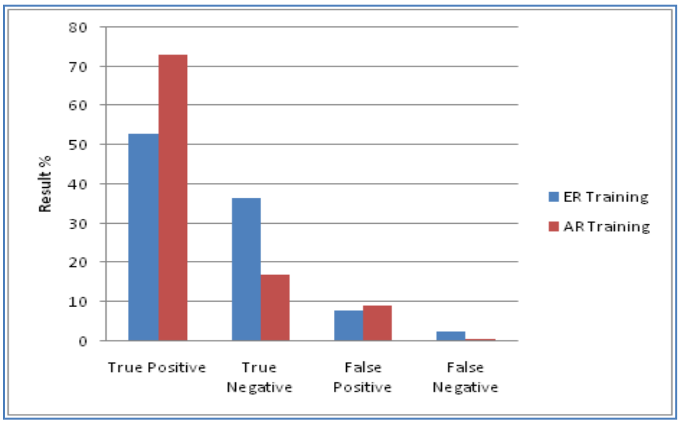

Fig-1 Predictive power of decision tree learning model over ER and AR training datasets

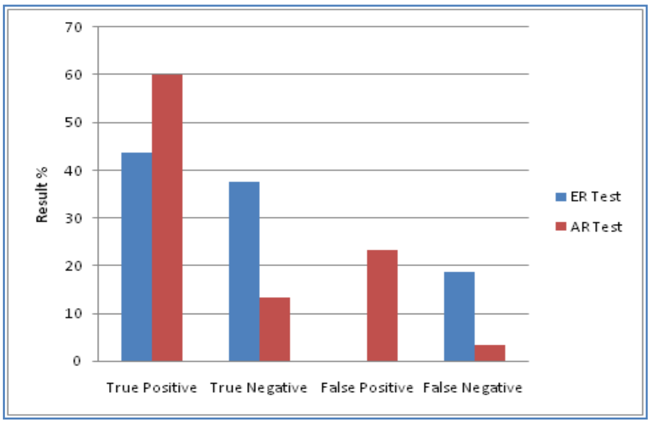

Fig-2 Predictive power of decision tree learning model over ER and AR test datasets

Table-2 *Actual & predicted logRBA values for test datasets*

Estrogen Receptor Dataset		Androgen Receptor Dataset	
Actual	Predicted	Actual	Predicted
-1.73	0.09	-0.79	-0.57
-3.07	-3.75	0.94	-1.03
-3.67	-3.18	-2.73	-2.79
-3.61	-2.20	-2.84	-2.67
-3.66	-3.18	-1.78	-2.56
-0.15	-0.51	0.07	-0.99
-0.67	3.06	2.05	0.55
-2.09	-2.70	-3.12	-2.38
-1.74	-2.45	-2.76	-2.53
-2.54	-3.67	-2.46	1.91
-3.22	-4.10	-3.46	-2.08
-3.22	-4.45	-1.61	-1.81
-3.44	-4.79	-1.64	-0.94
-2.74	-2.43	-0.35	-0.51
-2.82	-2.95	-2.05	-2.67
-0.19	-0.85	-3.17	-2.31
-0.69	-1.07	-2.74	-2.57
-2.30	-1.99	-2.27	-1.57
-3.16	-2.74	-0.74	-0.48
-0.35	-1.99	-1.98	-1.94

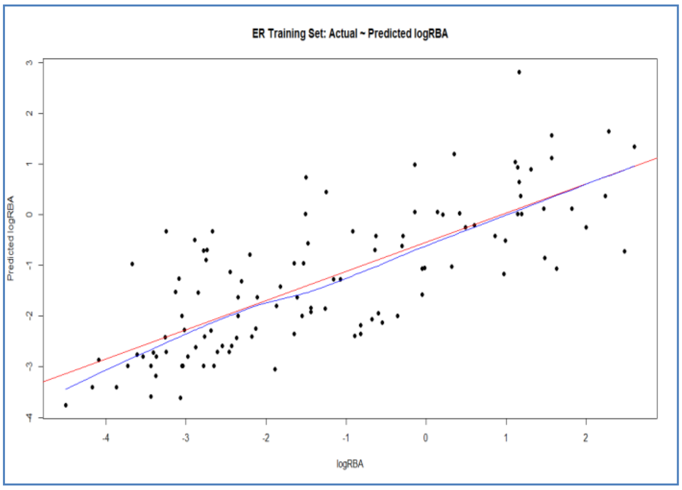

Fig-3 Actual versus predicted logRBA values resulted from ER training set linear regression model.
Red line -regression line (y~x), blue line - lowess line (x,y).

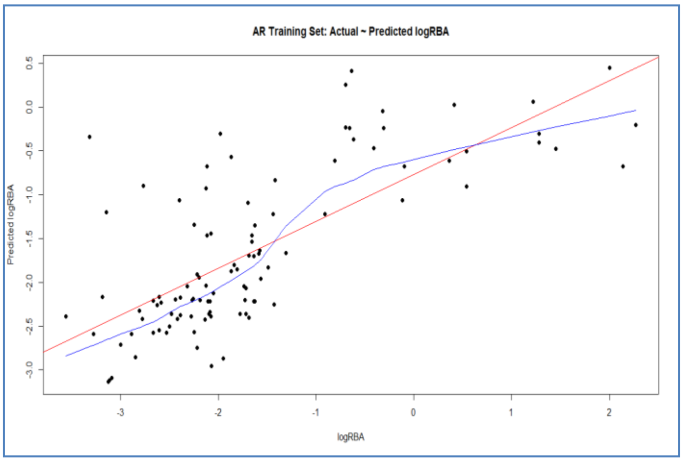

Fig-4 Actual versus predicted logRBA values resulted from AR training set linear regression model.
Red line -regression line (y~x), blue line - lowess line (x,y).

predicted value of AR and ER test datasets are given in [Table-2]. The F-statistic value of 6.911 for estrogen test set was greater than the table value [11] of 1.72 and the p-value of 1.156e^{-10} clearly reveals that this regression was statistically significant at the 95% confidence level and the R^2 value obtained was 0.5749, which state that the model can explain 57.49% accuracy. Similarly, for androgen test set the R^2 of the regression obtained was 0.5353, which means that the model can explain 53.53% of the accuracy in prediction. The F-statistic value of 3.548 for androgen test set was greater than the table value [11] of 1.65 and the p-value of 1.048e^{-05} expose that this regression was statistically significant at the 95% confidence level.

The proposed system is efficient similar to the earlier work [3] and it can be used to explore advance machine learning technologies for obtaining high accuracy in chemical toxicity prediction. This work can be expanded to more hormonal disruptions and other toxicity areas paving a way to develop an automated assistant tool to drug discovery scientists for helping them in lead chemical optimization for novel toxic free molecule discovery.

Conclusion

In this study, decision tree learning algorithm and linear regression algorithm were used for qualitatively and quantitatively predicting the endocrine disruption ability of the test compounds. Supervised classifier models were built with 200 estrogen disruptor and 160 androgen disruptor molecules as training datasets. Respective models were evaluated on 32 estrogen disruptor and 34 androgen disruptor molecules characterized with electro-topological state (E-state) fingerprints. Decision tree learning algorithm was found to be efficient on classifying chemicals as Endocrine disruptor or Non-Endocrine disruptor. Apache Spark tool on top of Hadoop Distributed File System (HDFS) was used to implement decision tree learning algorithm on parallel computing environment. Linear regression algorithm was applied over dataset in HDFS using R-Hadoop in R statistical software and logRBA values were predicted for disruptor chemicals. This work serves as example of predictive modeling in Cheminformatics on a big data platform. It can be further expanded for other hormonal disruptions and other toxicity areas so that an automated assistant tool for Drug discovery scientists can be developed for lead optimization to discover novel toxic free molecules.

Conflict Of Interest

The authors pronounce no competing financial interest

References

[1] Andrews N. (2008) *Lecture 1, Machine Learning*, Stanford, http://www.youtube.com/watch?v=UzxYlbK2c7E at 54 minutes (Accessed 6 December 2015).

[2] Collaborative *ltering.*http://www.csml.ucl.ac.uk/courses/msc_ml/?q=node/40 (Accessed 6 December 2015).
[3] Costantini L. and Nicolussi R. (2015) Performances Evaluation of a Novel Hadoop and Spark Based System of Image Retrieval for Huge Collections, *Advances in Multimedia*, http://dx.doi.org/10.1155/2015/629783 (Accessed 6 December 2015).
[4] Hester R.E. and Harrison R.M (Eds.), (2006) *Alternatives To Animal Testing (Issues in Environmental Science and Technology),* Royal Society of Chemistry, Cambridge.
[5] Mohan V. (2014) India bans import of cosmetics tested on animals, *The* Times *of India*. (Accessed 1 December 2015).
[6] Mukherjee R. (2014) Govt bans cosmetic companies from testing on animals, *The Times of India*. (Accessed 1 December 2015).
[7] Renjith P. and Jegatheesan K. (2015) *International Journal of Toxicological and Pharmacological Research,* 7(6), Article 8.
[8] Rokach L., Maimon O. (2008) *Data mining with decision trees: theory and applications,* World Scientific Pub Co Inc.
[9] Samuel L.A. (1959) *IBM Journal*, 3(3), 535-554.
[10] The *Apache Mahout Machine Learning Library*. [online] http://mahout.apache.org/ (Accessed 6 December 2015)
[11] Walpole, Meyers and Meyers, (1998) *Probability and Statistics for* Engineers *and Scientists*, 6th ed., Prentice Hall International Inc., New Jersey, 687-689.
[12] Wang L., Wang Y. and Xie Y. (2015) *Algorithms*, 8(3), 407-414.