Int J Pharm Bio Sci 2013 Apr; 4(2): (B) 1023 - 1029

Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences

ISSN 0975-6299

EFFICIENT MOLECULE REDUCTION FOR DRUG DESIGN BY INTELLIGENT SEARCH METHODS

A.KUMARAVEL*[1] AND PRADEEPA.R[2]

[1]Professor and Dean, Department of Computer Science and Engineering. Bharath University, Selaiyur, Chennai-600073, India,
[2]PG Student, Department of Computer Science and Engineering. Bharath University, Selaiyur, Chennai-600073, India,

ABSTRACT

Search methods applied to data mining techniques help us to analyze a data set. These methods are used to reduce the size of the search space in order to select the relevant molecules for the drug design. The research community in theoretical chemistry is very much depends on practical prediction and classification tools for this purpose. Classification is one of the major data mining methodologies. The objective of this paper is to check the learning algorithms for classification of drug design parameters based on 'Musk' qualifying dataset. The main intention in this context is to deal with a large data set with high accuracy. For this purpose Bayesnet, Naïve-Bayes, Decision table and random forest models are built using Weka tool under supervised learning algorithm. It is necessary to reduce the data dimension before constructing the models and thus the search methods for selection of attributes are followed. Those models are to be applied to predict the possible new drug candidates.

KEYWORDS: **Data mining, Classification, Drug design, Search Methods, Confusion matrix.**

A.KUMARAVEL
Professor and Dean, Department of Computer Science and Engineering.
Bharath University, Selaiyur, Chennai-600073, India,

1. INTRODUCTION

The fundamental goal is to predict whether the given molecule will bind to a target and it's how strongly during the process of drug design. The computational method is able to predict the biological activity of a molecule with high precision before a compound is synthesized. The biological activity of molecules is usually measured in analyzes to establish the level of inhibition of particular metabolic pathways. Drug discovery often involves the use of various structures to identify chemical structures that could have good inhibitory effects on specific targets. For example: activation of the oxygen atom in the molecule to a designated point in 3-space (OXY-DIS). In this paper we use data mining methods to improve computer-assisted drug design[11]. We apply four classification models based on different data mining approaches that can help us to select just a few molecules among more molecules. Classification methods belong to supervised learning. The models help us to identify molecules[9], which might be classified as biologically active. The parameters describe how effective the drug is. It indicates how much of a particular drug is needed to inhibit a given biological process by half. As the fig.1 illustrate our analysis belongs to the first stage of the process of finding new medications and it plays an important role in reducing the large data into a useable small subset. The next stages are the molecular scoring and docking, the methods of molecular classification (Bayes or Decision table or random forest[6] mechanical). The process ends with the synthesis and the laboratory testing of new drugs candidates.

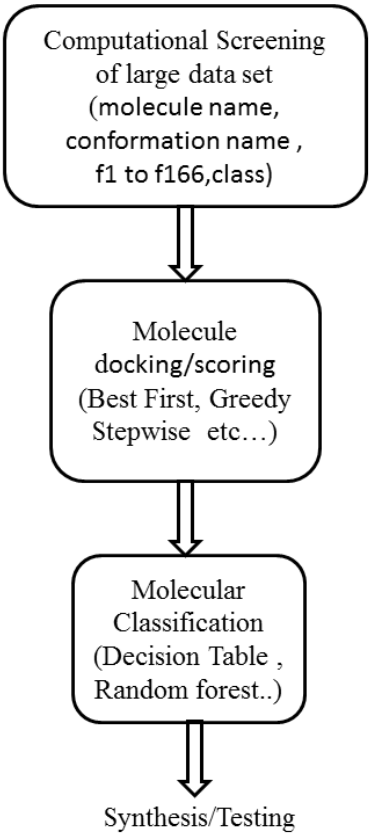

Figure 1
Process of new medications searching

2. EXPERIMENTAL ANALYSIS

In this section, we describe the data set and various search methods and compare the accuracy obtained with whole dataset and that with the selected attributes. Weka tool is used to reduce dimensionality with various select attribute techniques and classification models. The data sets for these experiments are from AI Group at 'Arris Pharmaceutical Corporation' donated by Tom Dietrich.

2.1 DATASET

2.1.1 DATASET DESCRIPTION

The database of MUSK "Clean2"is considered as our experimental dataset. It contains 39 musks and 63 non-musks, altogether 102 molecules. Number of Instances in this database is 6,598 and the number of attributes 168 plus the class[15].

2.1.2 ATTRIBUTE DESCRIPTION

The drug attributes are enumerated below with their description:

(1) *molecule_name:* Symbolic name of each molecule. Musks have names such as MUSK-188. Non-musks have names such as NON-MUSK-jp13.

(2) *conformation_name:* Symbolic name of each conformation. This attribute gives the combination of molecue number, stereoisomer number and confirmation number.This is denoted by MOL_ISO+CONF. The rest of the attributes denoted in the range f1 *through f162* which are "distance features" along rays. The distances are measured in hundredths of Angstroms. The value of distances may be negative or positive, since they are actually measured relative to an origin placed along each ray. The origin was defined by a "consensus musk" surface that is no longer used. The algorithm should not make any use of the zero point or the sign of each feature value[15]. *f163 through f166:* The distance of the oxygen atom in the molecule to a designated point in 3-space. This is also called OXY-DIS.OXY-X,OXY-Y and OXY-Z for x,y,z displacements from the designated points(f164,f165,f166). Finally the non-musk class is indicated by 0 and musk by 1. The attributes 'molecule_name' and 'conformation_name' are not used to predict the class as they are irrelevant and need not be considered for the attribute reduction.

2.2 METHODOLOGY

The first step of our analysis was to reduce the high data dimensionality. For this purpose we use Weka tool for attribute selection based on various search methods[16] made in the attribute space as shown in table1.We use factors which are selected after preprocessing as new predictors.

Table 1

The Elimination of Molecules by various search methods made in the attribute space.

S.no	Search methods	Cut off usage % for selection	Selected attributes for removal	Total no.of attributes removed
1.	Best first Search	0%	Conformation name, distance features (f1 to f35) & (f37 to f48)	47
		10%	Distance features(f91 to f95)	5
		20%	Distance features (f59,f64,f69,f73)	4
		30%	Oxy-X,Oxy-Y,Oxy-Z	3
2.	Greedy stepwise Search	0%	Distance features (f1 to f35) & (f37 to f48)	47
		10%	Distance features f13,f15, f91to f95	7
		20%	Distance features (f59,f64,f69,f73)	4
		30%	Oxy-X, Oxy-Y,Oxy-Z	3
3.	Ranker Search	0%	Distance features f16,f34,f37,f146,f157,f158,f159	7
		10%	Distance features f30,f60,f64,f138	4
		20%	-	-
		30%	-	-

2.2.1 METHOD DESCRIPTION

Best First Search

This method[12]searches the attribute subset space by best first search. The class for performing best first search method valid options are: (i) P option is the start set to specify a starting set of attributes, the attribute values are for example {1, 4, 7-9}. (ii) 'D' option is to specify the direction of search default is 1, where 0 denotes backward, 1 denotes forward and 2 denote bidirectional. (iii) 'N' option is to specify the number of non-improving nodes to consider before terminating search default is 5. (iv)'S' option is to specify the size of lookup cache for evaluated subsets.

Greedy Stepwise Algorithm

This method[13] performs a greedy forward or backward search through the space of attribute subsets. It may start with no attribute or all attributes or from an arbitrary point in the space. The process stops when the addition or deletion of any remaining attributes results in a decrease in the evaluation. It can also produce a ranked list of attributes by traversing the space from one side to the other and recording the order that attributes are selected.

Ranker

This method[14] performs ranking of attributes by their individual evaluations. It is used in conjunction with attribute evaluators (ReliefF, Gain Ratio, Entropy etc.).The method used in the ranker search is "get Num To Select()" to get the number of attributes to be retained. It's a kind of dummy search algorithm. It calls an attribute evaluator to evaluate each attribute not included in the start Set and then sorts them to produce a ranked list of attributes. The second step was to build different classification models on our training data set as a Bayesnet[7], Naive Bayes, Decision table and random forest models implemented in Weka tool. Analysis of various classification models and its accuracy of classifications are shown in table 2.

Table 2

Classification Method	Correctly classified	Incorrectly classified
bayes.BayesNet	92.09%	7.91%
bayes.NaiveBayes	85.78%	14.21%
rules.DecisionTable	99.75%	0.24%
trees.RandomForest	98.24%	1.75%

A confusion matrix illustrates the accuracy of the solution to a classification problem. Given m classes, a confusion matrix of dimension m x m which entry $C_{i,,j}$ indicates the number of tuples from D that were assigned to class C_j but where the correct class is C_i. The accuracy and error rates are calculated from confusion matrix[4],

Accuracy= (TP+TN) / (TP+FN+FP+TN) *100

Error rate= (FN+FP) / (TP+FN+FP+TN) *100

Where,

TP=True Positive

TN=True Negative

FP=False Positive

FN=False Negative

3. RESULT

The Best first model selected 10 molecules from data set and 9 molecules from normalized dataset. Greedy Stepwise selected 11 molecules from data set and 9 molecules from normalized dataset as

possible drugs table 3. The selected molecules are distance features (f13, f36, f49, f70, f76, f124, f126, f132) and distance of oxygen atom (f163).Our analysis continued by combining these results.

Table 3
No.of Attributes Selected

S.no	Methods	Before Normalization No.of attributes	After Normalization No.of attributes
1	Actual Data	168	168
2	BesatFirst-D 1-N 5	10	9
3	GreedyStepwise	11	9
4	Ranker	157	-

Quality of a certain model is often described by the confusion matrix. In this matrix each row represents the instances in an actual class; while each column represents the instances in a predicted class.The following tables shows the confusion matrix with error rates of classification models.

The error rate of the first model is 7.91%.
Bayesnet

Column1	Musk	Non-Musk
Musk	828	189
Non-Musk	333	5248
Error rate		7.91%

The error rate of the Second model is 14.21%.
Naviebayes

Column1	Musk	Non-Musk
Musk	783	234
Non-Musk	704	4877
Error rate		14.21%

The error rate of the Third model is 0.24%.
Decision Table

Column1	Musk	Non-Musk
Musk	1017	0
Non-Musk	16	5565
Error rate		0.24%

The error rate of the forth model is 1.75%.
Random forest

Column1	Musk	Non-Musk
Musk	903	114
Non-Musk	2	5579
Error rate		1.75%

Fig.2 shows the graph for accuracy of the various classification algorithms. DecisionTable method gives minimum error rate 0.24%.

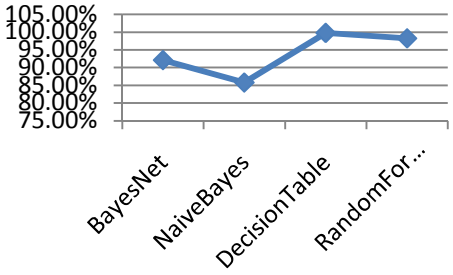

Figure 2
Comparison of main algorithms for accuracy

The results showed our subset of molecules can be selected. Moreover these molecules fulfill theoretical criteria to be new drug candidates.

4. CONCLUSION

The above results improve the previously obtained accuracies and the factors involved in the dimensionality of the search space. We have shown the search method i.e. Best first method reduces with an appreciable number of attributes and the rule based method i.e. Decision Table gives most accurate minimal error rate comparing the other methods. Therefore these models can be applied for prediction of other sets of molecules meant for further potential drugs candidates. It helps us to reduce the large data for exploring with usable smaller subset.

ACKNOWLEDGEMENTS

The authors would like to thank the management of Bharath University for the support and encouragement for this research work.

REFERENCES

1. https://www.waset.org/journals/waset/v68/v68-21.pdf world academy of science, engineering and technology, 2012.
2. Companion slides for the text by Dr. .H.Dunham, *Data Mining, Introductory and Advanced Topics*, Prentice Hall, 2002
3. Source about weka http://www.cs.waikato.ac.nz/ml/weka/ accessed on 14-01-2013.
4. http://webdocs.cs.ualberta.ca/~eisner/measures.html
5. A.Gelman, Y. S. Su, M.Yajima, J. Hill, M. Pittau, J. Kerman, and T. Zheng, "arm: Data Analysis Using Regression and Multilevel/Hierarchical Models," R package version 1.5-02.://CRAN.Rproject.org/package=arm,2012.
6. L. Breiman, " RandomForests,"in*Machine Learning*, vol. 45, pp. 5-32, 2001.
7. http://research.cs.queensu.ca/home/xiao/dm.html
8. Dietterich, T. G., Jain, A., Lathrop, R., Lozano-Perez, T. (1994). A comparison of dynamic reposing and tangent distance for drug activity prediction.Advances in Neural Information Processing Systems, 6. San Mateo, CA: Morgan Kaufmann. 216--223.
9. Jain, A. N., Dietterich, T. G., Lathrop, R. H.,Chapman, D., Critchlow, R. E., Bauer, B. E.,Webster, T. A.,Lozano-Perez, T.

Compass: A shape-based machine learning tool for drug design. Accepted for publication in Computer-Aided Molecular Design.
10. Dietterich, T. G., Lathrop, R. H., Lozano-Perez, T. Solving the multiple-instance problem with axis-parallel rectangles.
11. http://www.cs.princeton.edu/courses/archive /fall05/ cos597A/lectures/design.pdf, accessed on 15-01-2013.
12. http://weka.sourceforge.net/doc/weka/attributeSelection/BestFirst.html,accessed on 16-01-2013.
13. http://weka.sourceforge.net/doc.dev/weka/attributeSelection/GreedyStepwise.html, accessed on 16-01-2013.
14. http://weka.sourceforge.net/doc.dev/weka/attributeSelection/Ranker.html, accessed on 16-01-2013.
15. http://archive.ics.uci.edu/ml/datasets/Musk+%28Version+2%28, accessed on 16-01-2013.
16. Appavoo Kumaravel,' Clustering of different species based on mtDNA Sequences by frequent codons', Global Journal of Computational Intelligence Research-2012.