STUDY ON INFLAMMATION PROTEIN FOR ANTICIPATED COMPOUND ANALYSIS USING COMPUTAIONAL CLASSIFICATION ALGORITHM

M.Prasada Rao[1], Peri Srinivasa rao[2]

[1]*Dept of Computer Science and Systems Engineering, Andhra University, AP, India,* ***prasada.mandala@gmail.com***
[2]*Professor and HOD, Dept of Computer Science and Systems Engineering, Andhra University, AP, India,* ***peri.srinivasarao@yahoo.com***

Abstract

In machine language, as a field of empirical studies, the acquired expertise and knowledge from previous research has been guided the way of solving new tasks. The models should be reliable at identifying informative judgment and perceptive disease-treatment semantic relationships. These challenges are addressed by trying various predictive algorithms, and by using various textual representation techniques that we consider suitable for the task. As classification algorithms, in the present study we used decision based models using CART(Classification and Regression Tree) algorithm, out of six representative models which demonstrate possibilities offered by the Weka software to build classification models for protein ligand affinity analysis on COX2 protein with its novel compounds which can obtained from docking analysis and existed compounds respectively. In this case, the goal of classification models can be able to predict whether a new ligand will exhibit strong binding activity towards certain protein bio-targets according to their glide energy or distance or docking score which can be obtained from docking analysis using GLIDE software. In the concluding case one can expect that such ligands can have the corresponding type of biological activity and therefore could be used as hits for drug design.

Index Terms: *Classification Algorithm (CART), COX 2, Decision Tree, Inflammation, WEKA.*

--- *** ---

1. INTRODUCTION

Extensive usage of computers has made life easier for people in many aspects to carry out research and development in all fields. However, Computer science is widely used subject in the field of Bioinformatics and growing at a unique rate and the task in the early years of the millennium is to demonstrate how *in-silico* simulations [1] facilitate experiments in the laboratories and how this knowledge can be applied in curing human diseases. Due to this technological advancement, the amount of data that is generated in the world today had made decision making very complex. Computers have promised us a fountain of wisdom but delivered a flood of data [2]. Data mining is one approach that identifies the patterns in data and helps in making decisions by analysing this huge data ocean [3]. In this paper the work has been carried out on data mining techniques using Weka software.

The research experiments need to be guided such that high performance is obtained. The experimental settings are directed such that they are adapted to the domain of study medical knowledge and to the type of data we deal with short texts or sentences, allowing for the methods to bring improved performance. In this undergoing medical study the COX 2 protein and inhibitors affinity has been confirmed and classified. This Cox2 protein has role in inflammation- It is a part of the complex biological response of vascular tissues to harmful stimuli, such as pathogens, damaged cells, or irritants for a protective attempt by the organism to remove the injurious stimuli and to initiate the healing process. It is characteristically consideration of as a swelling, painful or otherwise uncomfortable situation where it possibly in your joints, sinus or intestine. Without inflammation, wounds and infections would never heal [4] [5][6].

There are at least two challenges that can be encountered while working with machine learning techniques. One is to find the most suitable model for prediction which can offers a suite of predictive models or algorithms that can be used and organized. The second one is to find a good data representation and to do feature engineering because features strongly influence the performance of the models. These challenges are addressed by trying various predictive algorithms and by using various textual representation techniques that we consider suitable for the task [7][8]. As classification algorithms, we use a set of six representative models: decision-based models or Decision trees, probabilistic models such as Naive Bayes and Complement Naive Bayes, which is adapted for text with imbalanced class distribution, adaptive learning, and a linear classifier [9]. We decided to use these classifiers because they are representative for the learning algorithms in the literature and were shown to work well on both short and long texts [10]. All classifiers are part of a tool called Weka [http://www.cs.waikato.ac.nz/ml/weka]. Two types of classified tasks will be considered as two class and multi class classification. In all cases protein-ligand binding data will be analyzed, ligands exhibiting strong binding affinity towards a certain protein being considered as active with respect to it. If it is not known about the binding affinity of a ligand towards the protein, such ligand is conventionally considered as non-active one.

2. MATERIAL AND METHODS

2.1 Data:

The following data has been taken from the Induced Fit Docking between the target protein 6COX and screened ligands 1-N-substituted-3, 5-diphenyl-2-pyrazoline derivatives were carried out using Glide software based on search algorithm [11],. The possible conformations of best ligands and native ligand along with their Docking score and Glide energy are showed in Table-1.

2.2 Weka:

It is a collection of machine learning algorithms for data mining tasks. It stands for the Waikato Environment for Knowledge Analysis, which was developed at the University of Waikato in New Zealand. The algorithms can either be applied directly to a dataset or called from your own Java code and contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization and runs on almost every platform. It is also well-suited for developing new machine learning schemes. It is used for research, education, and applications.

The easiest way to use it is through a graphical user interface called the Explorer. There are two other graphical user interfaces to Weka. The Knowledge Flow interface allows you to design configurations for streamed data processing. Another interface, the Experimenter, is designed to help you answer a basic practical question when applying classification and regression techniques: WEKA understands ARFF, CSV, C4.5 and binary file formats.

2.3 Data pre-processing and visualization:

Initial data was prepared as an input to Weka in the ARFF file format. First, it has to create in Excel file then save as filename.XLS, after that open the same file but now save as filename. CSV and save as type: CSV (delimited) then a file will be created. Again open CSV file with MS-Word and type the format of ARFF file as mentioned above, then save it as filename.ARFF and then as type : plain text. Finally, the ARFF file created and click on this file it will enter into WEKA environment for further process.

2.4 Cart Algorithm:

It stands for Classification and Regression Trees. It is a classical, statistical and machine learning technique was proposed by the Leo Breiman, Jerome Friedman, Richard Olsen and Charles Stone in 1984 [12]. Classification is a main job for in classified systematic allocation of objects is done according to their attributes. In classification problem, we have number of instances namely training data and predict which of several classes each instance belongs to. Each instance consists of several attributes, each of which takes on one of many possible values.

2.5 Classification:

In Weka explorer, by decision tree induction method the classification has been done. A decision tree is a flow-chart-like tree structure, where in each internal node denotes a test on the attribute, each branch represents an outcome of the test, and leaf nodes represents classes or class distributions. The top-most node in a tree is the root node. In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a leaf node that holds the class prediction for that sample. Decision trees can easily be converted to classification rules.

2.6 Decision Tree Induction in WEKA Explorer:

Open WEKA GUI, a window will open along with the menus, Click the explorer tab, a window with options then click on open file and load the desired file in WEKA. Now click the classify tab and click the choose button, to choose the desired algorithm and click on j48 algorithm. When click on start, the algorithm gets executed and the output will appear in the output window in the form of rules. To visualize tree, right click on the tree j48 under the result list and click on visualize tree.

For Classification by Decision Tree Induction in WEKA Knowledge Flow represents a graphical presentation which shows the threshold value to identify the effectiveness of the compound. For this analysis first

□ Open WEKA software, upon concerned window click on knowledge flow tab, then drag and drop the ARFF loader and load the desired data.

□ Now from the evaluation tab, drag and drop the class assigner icon and establish connection between ARFF loader and class assigner. This Class assigner is used to decide which attribute is considered as class attribute. For that right click on the class assigner and assign.

□ Now again from the same evaluation tab, drag and drop the cross-validation fold maker icon and establish connection between class assigner and cross validation fold maker. From the classifiers tab, drag and drop the j48 icon and establish connection between cross validation fold maker and j48 with training set as well as test set,

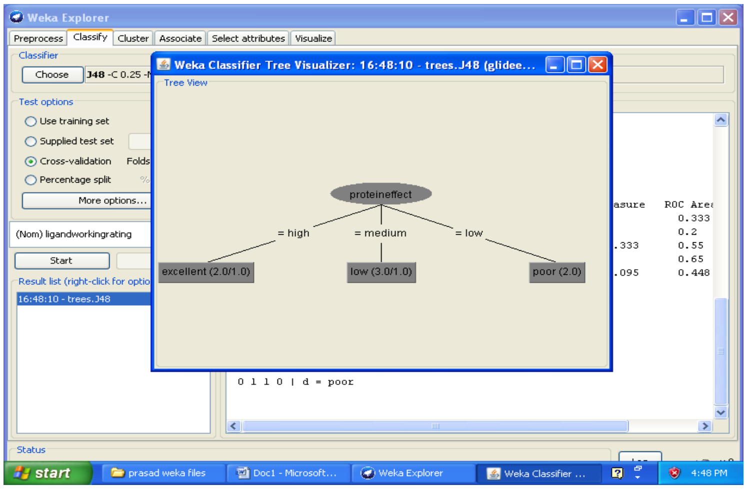
□ Now from the visualization tab, drag and drop the text viewer and graph viewer and establish connection with the j48. For running the algorithm, right click on the ARFF loader and click on start loading.

□ Then for viewing the results, right click on the text viewer and click on show results. Now for viewing the tree, right click on the graph viewer and click on show results. A small window by name graph viewer will come. When you click on the desired one, tree will get displayed

Table-1: Induced Fit Docking Results of Ligands against the Target 6COX

Compound	Distance (Å)	Docking Score	Glide energy Kcal/mol
Celecoxib	2.811	-11.45	-59.94
	3.364	-11.40	-55.85
	2.753	-11.11	-53.22
DuP – 697	2.765	-11.60	-61.66
	3.003	-9.98	-55.37
	2.679	-9.97	-52.89
Sc-58	3.256	-11.23	-59.85
	2.994	-11.34	-58.07
	2.794	-10.9	-59.15
Compound	2.849	-11.23	-62.71
	3.011	-10.78	-61.57
	3.237	-11.01	-58.95
Compound	2.925	-11.09	-61.13
	3.159	-11.05	-60.34
	2.860	-10.74	-59.82
Compound	2.73	-11.26	-62.79
	3.148	-9.10	-57.90
	3.156	-11.91	-57.36
Compound	3.030	-10.61	-61.17
	2.746	-10.61	-57.97
	2.859	-11.62	-55.15
Compound	2.98	-11.49	-63.00
	2.67	-11.04	-60.48
	3.14	-11.04	-58.77
Compound	2.89	-11.01	-61.72
	2.84	-10.80	-60.85
	2.94	-10.91	-58.91
Compound	3.16	-11.76	-64.11
	2.67	-11.29	-62.65
	2.66	-11.75	-64.02
Compound	2.75	-10.70	-64.08
	2.96	-9.32	-60.19
	2.917	-10.28	-59.88

3. RESULTS AND DISCUSSION


Weka Explorer

Classifier: J48 -C 0.25 -M

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds ○ Percentage split

(Nom) ligandworkingrating

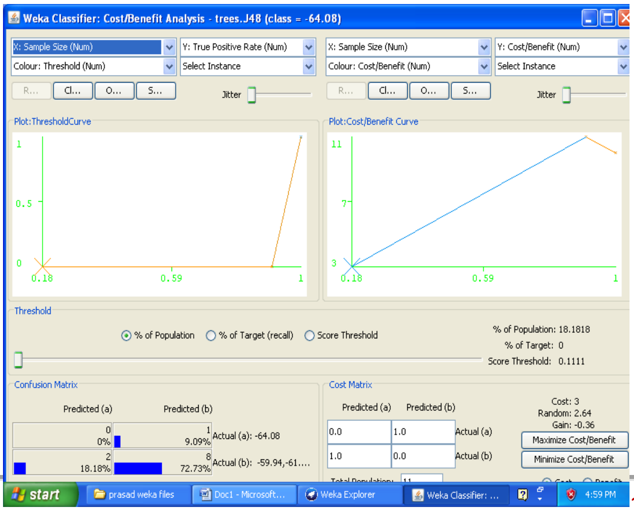
Result list (right-click for options): 16:48:10 - trees.J48

Weka Classifier Tree Visualizer: 16:48:10 - trees.J48 (glidee...

```
asure    ROC Area
          0.333
          0.2
.333      0.55
          0.65
.095      0.448

0 1 1 0 | d = poor
```

Fig1: Image viewing the desired output of decision tree using WEKA classification CART algorithm


Weka Classifier: Cost/Benefit Analysis - trees.J48 (class = -64.08)

X: Sample Size (Num) | Y: True Positive Rate (Num) | Colour: Threshold (Num) | Select Instance

X: Sample Size (Num) | Y: Cost/Benefit (Num) | Colour: Cost/Benefit (Num) | Select Instance

Threshold: ● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 18.1818
% of Target: 0
Score Threshold: 0.1111

Confusion Matrix

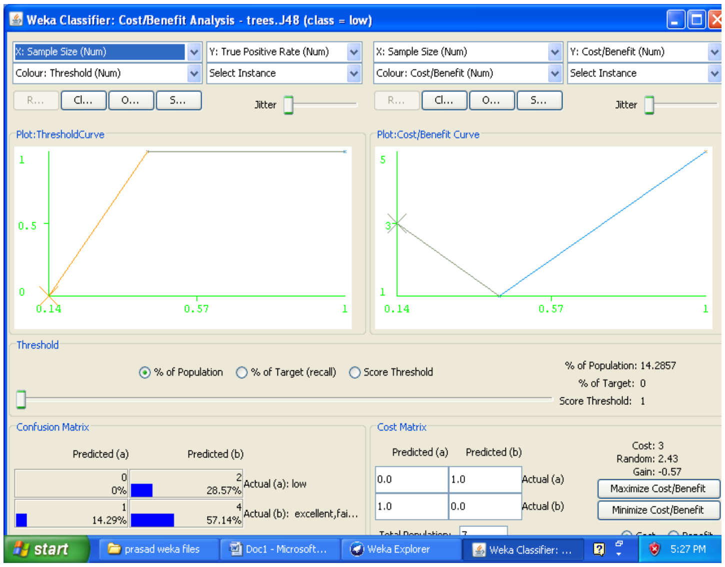
	Predicted (a)	Predicted (b)	
	0 0%	1 9.09%	Actual (a): -64.08
	2 18.18%	8 72.73%	Actual (b): -59.94,-61....

Cost Matrix

Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Cost: 3
Random: 2.64
Gain: -0.36

Fig-2: Graphical representation of compound 12 (proposed) with its threshold value via WEKA tool



Weka Classifier: Cost/Benefit Analysis - trees.J48 (class = low)

X: Sample Size (Num) | Y: True Positive Rate (Num) | Colour: Threshold (Num)

X: Sample Size (Num) | Y: Cost/Benefit (Num) | Colour: Cost/Benefit (Num)

Threshold: ● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 14.2857
% of Target: 0
Score Threshold: 1

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	2 28.57%	Actual (a): low
1 14.29%	4 57.14%	Actual (b): excellent,fai...

Cost Matrix

Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Cost: 3
Random: 2.43
Gain: -0.57

Fig 3: Graph representing the threshold curve of compound3(existed one) with its threshold value

Based on the threshold value graph in Figure-2 described about compound12 is shown above.

The threshold value of compound12 is less than 1 so it is the effective compound when compared to the others.

If the value of protein is reached to the value 1 in the graph in Figure-3 and that reveals the threshold value is high. So that, the protein has low energy. When compared with studied docking analysis, it is not effective. If the value of the protein is not reached the value of 1 then the threshold value of the protein is low. Then the protein is effective to compare with other protein glide energy values. According to above graphs, when the comparison among different proteins and its values which have been obtained from docking studies using GLIDE software in structure based design. The protein whichever having high glide energy have low threshold value and the proteins whichever having low glide energy have high threshold value. As per the graphical representation using WEKA tool and bioinformatics docking tool GLIDE both have the same results. Therefore, the analogous results showed here have been explicitly noted that both the integrated technologies would have the similar effects.

3. CONCLUSION

Cox 2 is a striking and prospective target for inflammation. In recent years, potential applications for COX 2 inhibitors have emerged in the treatment. In the present undergone study based on the molecular docking results obtained from novel ligands of 1-N-substituted-3, 5-diphenyl-2-pyrazoline derivatives along with existed drugs by search algorithm using Glide software showed that the compound 12 has best docking score (-11.76) and Glide energy (-64.11) compared to that of the existing drugs. Using the classification technique based on CART algorithm by tool WEKA; by means of this software, the glide results have been studied and identified the decision tree the threshold values calculated. Based on this calculated values the protein whichever having high glide energy have low threshold value and the proteins whichever having low glide energy have high threshold value. As per the graphical representation using WEKA tool and bioinformatics docking tool GLIDE both have the same results. Hence, it depicts and states that the analogous results have been explicitly confirmed the classified algorithms and search algorithms. It is noted that both the integrated technologies were presented in this thesis would be a quite definite step to advance in this field.

REFERENCES

[1]. Sieburg, H.B. Physiological Studies in silico. Studies in the Sciences of Complexity 1990 12: 321–342.

[2]. William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus.Knowledge Discovery in Databases: An Overview.AI Magazine 1993 Volume 13 Number 3

[3]. Witten, Ian H.; Frank, Eibe; Hall, Mark A. *Data Mining: Practical Machine Learning Tools and Techniques* (3 ed.). 2011 Elsevier. ISBN 978-0-12-374856-0.

[4]. Serhan CN, Savill J. Resolution of inflammation: the beginning programs the end. *Nat. Immunol.* 2005 **6**(12): 1191–1197. PMID 16369558.

[5]. TShimokawa et al J. Bio. Chem. 1990 265 (33): 20073-20076 [PMID: 2122967].

[6]. Vane JR,Nat New Biol. 1971 Jun 23;231(25):232-5.[PMID: 528436]

[7]. I. Donaldson et al BMC Bioinformatics. 2003 Mar 27; 4:11. Epub 2003 Mar 27 [PMID: 12689350]

[8]. R. Gaizauskas et al Bioinformatics. 2003 Jan; 19(1):135-43.[PMID: 12499303]

[9]. A.M. Cohen and W.R. Hersh, and R.T. Bhupatiraju, Feature Generation, Feature Selection, Classifiers, and Conceptual Drift for Biomedical Document Triage, Proc. 13th Text Retrieval Conf.(TREC), 2004.

[10]. R. Bunescu, R. Mooney, Y. Weiss, B. Scho¨ lkopf, and J. Platt.Subsequence Kernels for Relation

Extraction.Advances in Neural Information Processing Systems, 2006 vol. 18, pp. 171-178.

[11]. M PrasadaRao, P SrinivasaRao, AllamAppaRao and Manda Rama NarasingaRao, computer aided design and molecular docking study of 1-n-substituted-3, 5-diphenyl-2-pyrazoline derivatives as cox–2 inhibitors, Int. J. Bioassays, 2013 02 (11), 1453-1456.

[12]. Breiman L., Friedman J., Olshen R., and Stone C. Classification and Regression Trees. Wadsworth Int. Group, 1984

BIOGRAPHIES

Description about the author1

Mr. Prasada Rao Mandala, working as Associate Professor in Pydah College of Engineering and Technology, Visakhapatnam, AP, India. Over 10 years experience in teaching and administration. Published papers in so many international journals and conferences. Organized and participated in many international, national workshops and conferences.

Description about the author2

Prof. Peri Srinivasa Rao, working as a Professor and Head of the Dept of AUCSSE. Andhra University, Visakhapatnam, AP, India.
Papers in journals: 26, No of PhDs guided: 6.