Arrow's Ensemble Feature Selection Machine Learning Framework for Drug Toxicity Prediction

Nishtha Hooda
Computer Science and Engineering Department
Chandigarh University
Mohali, Punjab
27nishtha@gmail.com

***Abstract*—Big Data predictive analytics using machine learning techniques is currently much active area of research in biological computing. With increasing size and complexity of drug molecular data during drug designing and experimentation, deep learning gained huge success in harnessing high dimensional molecular descriptors. In this research work, DCNN (deep convolutional neural networks) an efficient prediction model, combined with arrow theorem based feature selection algorithmic characteristics is proposed, which can predict whether a drug is toxic or not. The experiments are carried out on high dimensional drug data and achieved an accuracy of 83% and AUC of 0.63. Promising results are found, when the performance of the proposed framework is compared with the standard classifiers like SVM, random forest, etc. using different evaluation metrics like accuracy, sensitivity, etc. With the appearance of increasing impact of toxic drugs in human life, deep learning will play a big part in improving the quality of prediction of drug toxicity in the future.**

Keywords—big data; machine learning; prediction; deep learning; drug toxicity

I. INTRODUCTION

Identifying and monitoring toxic drug molecules for human safety are key challenges in the area of drug designing and development [1]. Deep learning and Big Data are two hottest trends in the rapidly growing digital world [2]. While Big Data has numerous definitions, this research work refer it to the variety i.e. unstructured drug molecular descriptors data as presented in the Figure 1, defining important Vs of big data [3].

Classification models can be employed to detect drug toxicity based on molecular-descriptors of drug molecule samples. The classifiers are trained using data collected from various drug molecule samples. Government agencies, NIH, EPA etc. instigated the Tox21 data challenge to encourage the ingenious computational techniques for drug toxicity prediction [4]. The drug data is complex and difficult to analyze using conventional data analysis techniques. Hence, deep learning offers a great solutions in harvesting valuable knowledge from such complex medical data.

For high dimensional nature drug data, researchers and practitioners are putting forth hard efforts to improve the accuracy of classification model through innovative feature selection techniques [5]. Plenty of feature selection techniques are available to remove extraneous features from the original high dimensional data set [6, 7, 8]. Random subset selection is predominantly implemented by researchers for feature selection [9, 10]. Theoretical studies show that ensemble ensemble of feature selection yields more accurate results [11, 12, 13]. For this purpose, instead of using the results of one feature selection approach, the admired technique is to implement diversity and combine the results of several feature selection techniques. For instance, the diversity is achieved by combing the results of different feature selection rankers using aggregation techniques like arithmetic mean, majority voting [11, 12, 13].

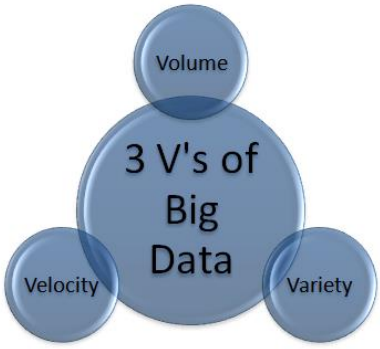


Figure 1 3 V's of Big Data [3]

Although Arrow's impossibility theory axioms has been considerably used by mathematicians and practitioners in the voting theory [14, 15]. It is implicated in engineering design decision making by many research works, but it has rarely been connected specifically in modeling the machine learning modeling problem [16, 17, 18, 19, 20]. Connecting it for ensemble feature selection, validity of arrow's theory axioms and constraints are studied for collective preference of several feature rankers.

The main intent of defining axioms of Arrow's theorem for an ensemble feature selection is to model the machine learning problem mathematically. Besides modeling the problem mathematically, checking the validity of these axioms for the machine learning real world problem helps in drawing fair and satisfactory conclusions at the end. Using an Arrow's axiomatic aggregation A^2EFS framework for an ensemble feature selection for improving the performance of the designed model.

Due to the complex and high dimensional drug data, conventional methods encounters nasty repercussion. In this paper, we proposed an Arrow's Axiomatic based Ensemble Feature Selection (A^2EFS) framework for drug toxicity classification problem. The performance of developed framework is also compared with the standard state-of-the-art classifiers like SVM, random forest, bagging, etc even when they are combined with classical feature rankers.

The rest of the paper is organized as follows: Section 2 briefly describes the different methods like class-balancing, feature selection, etc. that are used in the framework. It also discusses the motivation, Arrow's constraints and an algorithm for implication of ensemble feature selection method. Section 3 discusses the data, its features along with proposed framework and experimental setup. Section 4 summarizes the experimental results and finally, Section 5 discusses the conclusion and future scope.

II. MATERIAL AND METHODS

The main intent of research in machine learning is to create the programmed computing machine for the classification and prediction.

Motivation

Arrow's research work is used in the context of social welfare economics [14, 15]. Political researchers have to work on finding the fair method for combining the preferences of potentially conflicting parties or set of people in the society. Majority voting is one such fairest method for fair decision making. The economists desire to choose the most satisfactory alternatives from the society. Similarly the intent of machine learning framework with high dimensional data set problem is to train the model with the best set of features. Rationale of the two problems is to select the best subset of candidates from a huge set of nominees using some fair and satisfactory system which gives an evident that both the problems can be formalized in the similar manner indeed.

The designed A^2EFS framework has adopted an ensemble technique to extract the best features of drug data that are contributing to the classifier's performance. One efficacious approach for generating an ensemble is to utilize the work of multiple feature selection techniques. While the goal of classical method of feature selection is to find only the best feature-subset pertinent to learning task, an ensemble feature selection focuses on the diversity and producing the feature subset by preserving the agreement among all the base classifiers.

To create an ensemble feature selection, we have focused on two important aspects:

- Which feature selection techniques are used to rank the drug molecule features?
- How to integrate the ranking of different techniques to introduce fair ranking system.

The integration technique which is generally used by the researchers is the arithmetic mean method. This method defines score to the features in the ranking list and then final score of the features are summed up to get the final ordered list. Highly ranked features of the training data are then selected from the list of these final ordered list. Although there are many other ways to integrate the individual feature ranking choices, the performance gain may depend on the procedure by which final collective ranking is performed. To study this, we need to tactfully model the individual's rankings and explore the criteria by which aggregation methods can be compared. Condorcet's Paradox of voting explored many aggregation procedures and compare their properties [21]. In 1950, the famous impossibility theorem for social choice was presented by Kenneth Arrow. Arrow's original work was proved on the problem of aggregating the preference of different individuals into common order as a whole. Different constraints and limits posed by Arrow's theorem in aggregating a different set of orders into a single one are studied and adopted in the designed A^2EFS framework.

Arrow's Axioms for Aggregation

Arrow's theorem also proved that as the number of options to choose from is equal to or greater than three, no aggregation method for collective ordering can satisfy all these desiderata [22]. Connecting it to machine learning and feature ensemble aggregation problem, different feature selection rankers have different preference order depending upon the criteria feature-rankers are ranking the set of features. The aggregation procedure is expected to have following properties or constraints which are expressed by Arrow's desiderata [22]:

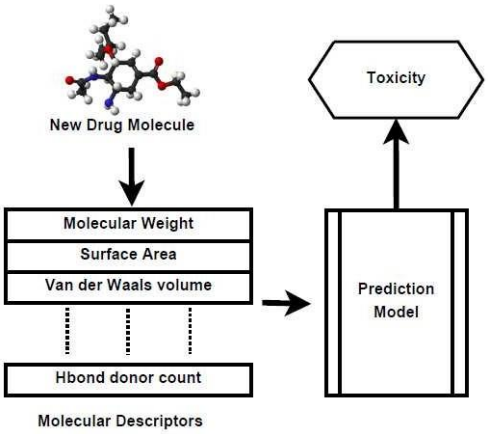
- *Unrestricted domain* (Axiom 1): It reflects universality which defines that the chosen aggregation method should be ready for all types of individual preferences by all the voters. It should also yield a unique and complete ranking of all the available choices.
- *Consistency and rationality* (Axiom 2): It is an ordinal version of monotonicity. The collective

choice function should combine the different preference orders in a sense that the collective preference order as a whole on a set of options must be connected as well as transitive.

- *Independence of irrelevant attributes* (Axiom 3) : It defines that when the collective preferences are assembled then, it should mirror one pair of alternatives at a time.
- *Not imposed* (Axiom 4) : An order is said to be imposed if some feature X is ranked before feature Y in all the profiles. This states that the preference of any feature ranker should not be excluded from being chosen in the final order.
- *Non-dictatorship* (Axiom 5): It demands that the results of one voter cannot be so privileged that their preference order dictates the final order of features.

Proposed Algorithm

The goal of the research is to design and develop a prediction model for drug toxicity prediction. The outcome of A^{2}EFS Framework will be available in the form web-based application that helps to predict the activity of the newly discovered chemical compound as described as abstract view in the Figure 1. The prediction method is described in the Algorithm 1.



A2EFS Framework can also be used as a decision support system (described in Figure 3) that helps to improve the activity of a drug molecule. For simplicity, the complex drug data is processed and sub-sampled to consider it as a binary-classification problem. Let (F1, F2, F3..Fn) represents the behavioural data of 200000 samples fitted into a model (Wi, Xiǫ{c0, c1}), where c0 is the class of non-toxic drug molecules and c1 is the class of toxic molecules. Now n being very large, selecting optimal features subset for the prediction model and handling class balancing is indispensable part of the study and described in Figure 2.

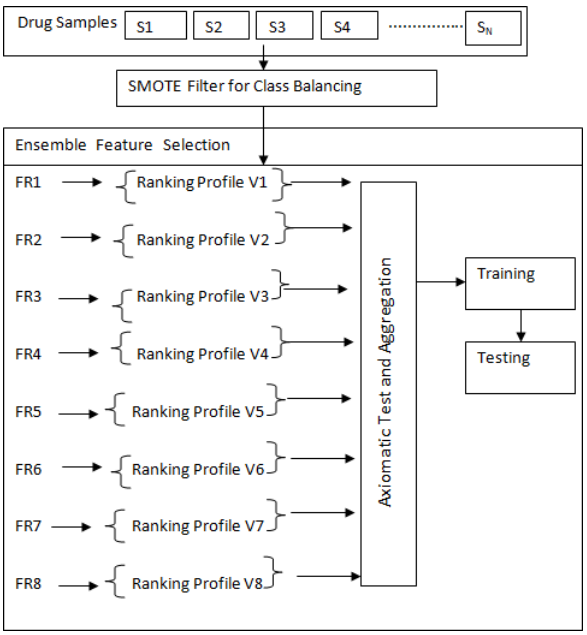


Figure 2 Proposed framework

Majority voting is one such fairest method for fair decision making. The intent of machine learning framework with high dimensional data set problem is to reduce the complexity and to train the model with the best set of features.

Rationale of the problem is to select the best subset of candidates from a huge set of nominees using some fair and satisfactory system The designed framework has adopted an ensemble technique to extract the best features of drug data that are contributing to the classifier's performance. One efficacious approach for generating an ensemble is to utilize the work of multiple feature selection techniques as shown in the Figure 1.

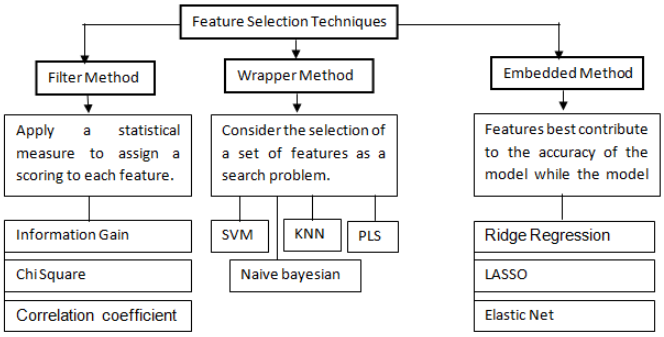


Figure 3 Ensemble feature selection

While the goal of classical method of feature selection is to find only the best feature-subset pertinent to learning task, an

ensemble feature selection focuses on the diversity and producing the feature subset by preserving the agreement among all the base classifiers. [19, 20].

III. EXPERIMENTAL INVESTIGATION

This section discusses about the dataset and experimental setup.

A. Dataset

Data is collected from TOX21 data challenge.

B. Experimental Setting

The R 'caret' package is used to implement the various feature selection techniques. Each of the steps described in the Figure 1 are performed on the data set. Purpose is to measure the prediction accuracy of the classifier of new samples without the benefit of knowing the true class of the samples. To test the robustness of designed A^2EFS framework, the process is iterated. To evaluate the performance of the proposed A^2EFS framework, ten different parameters namely accuracy, kappa, sensitivity, specificity, prevalence, positive prediction value, negative prediction value, detection rate, detection prevalence and area under curve (AUC) are used.

IV. RESULTS AND DISCUSSION

This section discusses parameter evaluation metrics to measure the performance of various machine learning algorithms. The result is represented graphically.

A. Performance Evaluation

The performance of the proposed framework is evaluated with different parameters of confusion matrix as shown in the Table 1. The framework performance is evaluated with different performance parameters of confusion matrix as shown in the Table 1 and 2 are used to evaluate and summarize the results of a proposed framework.

The performance of ensemble feature selection method is compared with the outstanding state-of-the-art classifiers like SVM, random-forest, C5.0, glm, knn, pls, adabag, etc. even when they are using a single feature selection technique. The performance results of these classifiers when they are using ensemble feature selection are described in the Table 3.

As accuracy is not the best metric in case of imbalanced data [21, 16, 18], results in the Table 3 have summarized different other metrics like sensitivity, specificity, etc. Table 4 summarizes the result of accuracy of different classifiers when they are combined with different feature rankers like rf, pls, glm, etc. individually. The last column of the table is highlighted to show the improvement in accuracy when the same classifiers are executed with ensemble feature selection method. It is also represented graphically in the Figure 5. Similarly in Table 5, we have observed that value of AUC using ensembling feature selection is improved in every classification method, as represented in the Figure 5.

Table 1 Confusion Matrix

	True Reference	
Predicted Condition	Condition Positive	Condition Negative
Positive	True Positive X	False Positive Z
Negative	False Negative Q	True Negative Y

Table 2 Comparing performance of classifiers

Parameters	Accuracy (%)	Sensitivity	Specificity	AUC
SVM	80.02	0.919	0.677	0.798
Glmnet	81.20	0.70	0.896	0.798
Glm	75.60	0.709	0.792	0.75
Knn	66.10	0.653	0.673	0.67
Rf	86.70	0.806	0.75	0.819
Pls	55.80	0.522	0.657	0.672
Adabag	76.40	0.786	0.747	0.763
C5.0	78.20	0.823	0.739	0.781

Table 3 AUC Comparison of various feature selection methods

Models	Rf	Pls	Lda	gbm	rknn	Knnn	glm	glmnet	ensemble
SVM	0.71	0.72	0.60	0.69	0.65	0.73	0.71	0.62	0.80
Glmnet	0.77	0.70	0.72	0.73	0.66	0.74	0.74	0.76	0.80
Glm	0.61	0.78	0.66	0.70	0.73	0.58	0.75	0.56	0.75
KNN	0.63	0.65	0.63	0.64	0.65	0.62	0.60	0.64	0.67
RF	0.78	0.78	0.81	0.78	0.84	0.86	0.84	0.80	0.82
PLS	0.55	0.53	0.55	0.56	0.62	0.50	0.66	0.51	0.67
Adabag	0.73	0.68	0.76	0.69	0.67	0.72	0.69	0.76	0.76
C5.0	0.73	0.72	0.76	0.76	0.76	0.75	0.81	0.79	0.78

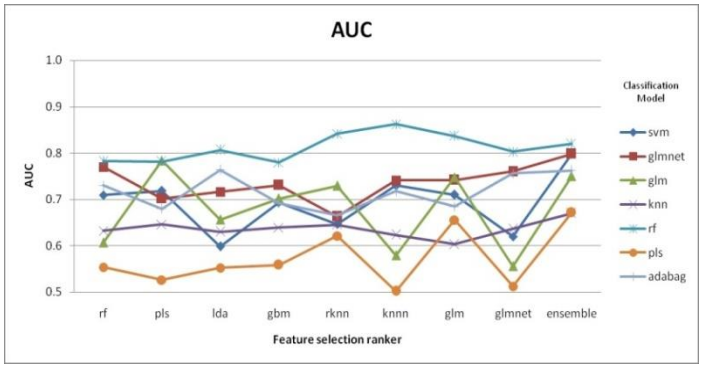


Figure 4 Comparing AUC with ensemble

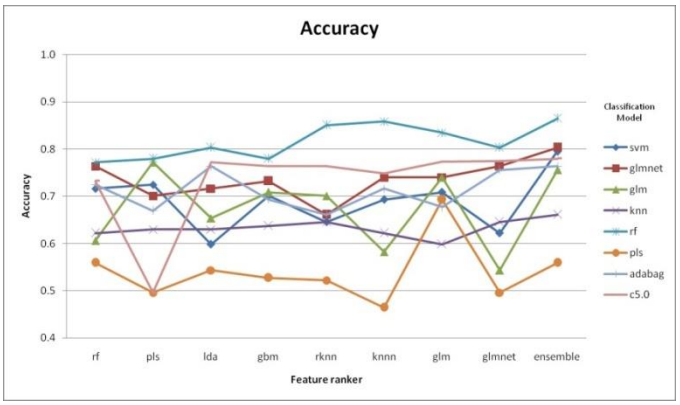


Figure 5 Comparing accuracy with ensemble

V. CONCLUSION

In this paper, we proposed a framework called A^{2}EFS framework to solve the classification and pre- diction task for drug toxicity assessment. As the collected data is highly imbalanced and suffers from the curse of dimensionality, SMOTE filters and an innovative ensemble feature selection are implemented in the framework. Arrow's impossibility theorem is presented and it's axioms are implicated for modeling an ensemble feature selection problem mathematically. Validity of Arrow's axioms and constraints are studied for collective preference of several feature rankers. Testing the validity of these axioms for aggregating the results of several feature rankers helped in drawing fair and satisfactory results. The proposed framework managed to produce a promising results regardless of high-dimensionality and class-imbalance problems. Extensive comparison of the proposed feature selection method with other feature rankers by combining them with different state-of-the-art classifiers with different goodness-of-fit classification metrics serves as proof of eligibility of proposed framework to perform an efficient toxicity assessment.

For future works, we are suggesting to enhance the A^{2}EFS framework by implementing it on the top of modern big data techniques like Hadoop, Spark,etc.

References

[1] T. G. Neltner, H. M. Alger, J. E. Leonard, M. V. Maffini, Data gaps in toxicity testing of chemicals allowed in food in the united states, Reproductive Toxicology 42 (2013) 85–94.
[2] Chen, Xue-Wen, and Xiaotong Lin. "Big data deep learning: challenges and perspectives." IEEE access 2 (2014): 514-525.
[3] Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." International Journal of Information Management 35.2 (2015): 137-144.
[4] Hooda, Nishtha et al. "B2FSE framework for high dimensional imbalanced data: A case study for drug toxicity prediction." Neurocomputing 276 (2018): 31-41.
[5] C. Chu, A.-L. Hsu, K.-H. Chou, P. Bandettini, C. Lin, A. D. N. Initiative, et al., Does feature selection improve classification accuracy? impact of sample size and feature selection on classification using anatomical magnetic resonance images, Neuroimage 60 (1) (2012) 59–70.
[6] V. Bol´on-Canedo, N. Sa´nchez-Maron˜o, A. Alonso-Betanzos, A review of feature selection methods on synthetic data, Knowledge and information systems 34 (3) (2013) 483–519.
[7] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Computers & Electrical Engineering 40 (1) (2014) 16–28.
[8] N. Dess`ı, B. Pes, Similarity of feature selection methods: An empirical study across data intensive classification tasks, Expert Systems with Applications 42 (10) (2015) 4632–4642.
[9] O. Ra¨s¨anen, J. Pohjalainen, Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech., in: INTERSPEECH, 2013, pp. 210–214.
[10] Q. Song, J. Ni, G. Wang, A fast clustering-based feature subset selection algorithm for high-dimensional data, Knowledge and Data Engineering, IEEE Transactions on 25 (1) (2013) 1–14.
[11] Y. Saeys, T. Abeel, Y. Van de Peer, Robust feature selection using ensemble feature selection techniques, in: Machine learning and knowledge discovery in databases, Springer, 2008, pp. 313–325.
[12] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics 26 (3) (2010) 392–398.
[13] F. Maniquet, P. Mongin, Approval voting and arrows impossibility theorem, Social Choice and Welfare 44 (3) (2015) 519–532.
[14] B.-G. Ju, Individual powers and social consent: An axiomatic approach, Social Choice and Welfare 34 (4) (2010) 571–596.
[15] Y. Saeys, T. Abeel, Y. de Peer, Towards robust feature selection techniques, in: Proceedings of Bene- learn, Citeseer, 2008, pp. 45–46.
[16]M. J. Scott, E. K. Antonsson, Arrow's theorem and engineering design decision making, Research in Engineering Design 11 (4) (1999) 218–228.
[17] G. A. Hazelrigg, The implications of arrows impossibility theorem on approaches to optimal engineering design, Journal of Mechanical Design 118 (2) (1996) 161–164.
[18] M. Franssen, Arrows theorem, multi-criteria decision problems and multi-attribute preferences in engi- neering design, Research in engineering design 16 (1-2) (2005) 42–56.
[19] G. A. Hazelrigg, Validation of engineering design alternative selection methods, Engineering Optimiza- tion 35 (2) (2003) 103–120.
[20] G. A. Hazelrigg, A framework for decision-based engineering design, Journal of mechanical design 120 (4) (1998) 653–658.
[21] S. J. Brams, P. C. Fishburn, Voting procedures, Handbook of social choice and welfare 1 (2002) 173–236.
[22] K. J. Arrow, A difficulty in the concept of social welfare, The Journal of Political Economy (1950) 328–346.