Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Application of Neural Network in Virtual Screening of Bioassay

Prajna Manoor Kumar [1]
Computer Science, University of Central Florida, Orlando, Florida, 32816, USA

Abstract

Motivation: Virtual Screening of bioassays associates with three main problems. This paper mainly focusses on tackling those main problems. First, data in bioassay compounds are highly imbalanced, data usually has the low ratio of Active compounds to Inactive compounds. Secondly it is very difficult to get curated data. Thirdly the data has large number of false positives that occurs in the physical primary screening. This paper discusses about the application of the neural network as the solution in alternative to solution proposed by the previous work, that is the application of the Weka cost sensitive classifiers (Naïve Bayes, SVN, C4.5, Random Forest).

Results: In comparison to Weka cost sensitive classifiers neural network outperforms these classifiers in terms of accuracy of ability to hold large number of data. Use of confirmatory screened data alone improved the quality of dataset that resulted in lower rate of misclassification.

Availability: Virtual Screening of bioassay using Weka cost sensitive classifiers.

Contact: Journals of cheminformatics.

Supplementary information: Supplementary data is available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2820499/.

1 Introduction

Bioassay is the process of determining the potency of a drug and its effect against the biological targets. Developing new drugs for a disease is time consuming and expensive process. The drugs have to undergo a various kind of tests for several years and over several millions will be spent just to bring the drug to market. This development of new drug involves it to subject to High Throughput Screening (HTS). It is the combination of the robotics, control software, liquid handlers and optical readers. In this process drug under observation is tested against the various biological compounds to check if the drug can bind to the required target. If the drugs bind to the target, then it is hit or else it is a miss. Through this method we can identify the active compounds. Genes or molecule that modulate a biomolecular reaction. The result of these experiments gives a starting point for further drug design and understand how a works on a target.

Virtual Screening is the in-silico screening of the biological samples and it complements High Throughput Screening in narrow downing the drug selection. Virtual Screening is largely been a number game focusing on how a humongous chemical library be filtered to manageable chemical numbers so that it can be synthesized purchased and tested. Today Virtual Screening has become an integral part of the drug discovery. Virtual Screening utilizes several computational techniques that are available today based on the type and amount of the information available about the compound that is being tested. If the available active compound is quite less in number, then techniques like structure similarity is used. If the number of compounds is good in number, then other analysis techniques like machine learning and deep learning can be applied. Selecting a compound whose activity is known then classifying active or inactive compounds by building a predictive model. Ultimately using this knowledge to screen the other unknown compounds so that it is most likely to be active for a selected screening.

The major challenge involved when using a machine learning was imbalanced data. Any tool that is employed in Virtual Screening should be able to cope up with this highly imbalanced data. Virtual Screening of pharmaceutical imbalanced data has been carried out before as well. This paper we will be talking about the application of the neural network that deals with the problem of the misclassification due to imbalanced data. Some of the previously done works involves the application of the Naïve Bayes to the PubChem Bioassay that reduces the active and inactive compound ratio to 1:1 and then proceeds with the classification and they did not take misclassification cost into consideration. The focus is on the confirmatory dataset than keeping both primary and
confirmatory screening.

2 Methods

3.1 Biodata Collection

The dataset for this project is the Bioassay that is used in the screening process of a drug. This data can be collected from and database system called PubChem, a database system for chemical molecules. It contains information regarding various chemicals and their activities against the biological samples. This database system is maintained by the National Center for Biotechnology Information (NCBI), it us under the National Library of Medicine, which intern is a part of United States National Institutes of Health. This database system is an open database system which means one can input the AID number of a bioassay and download the data table. Variety of dataset has been chosen for this study. The dataset used are the dataset obtained from different screening which can be used in High Throughput Screening technology. Table 1. Gives brief summary over the dataset used in this experiment. In this project author has used only the confirmatory data for a particular type of a target. There were more than one confirmatory dataset and they were joined in the further process. This is done to check if the merged data gives better result than single dataset alone.

Table 1. Summary of Bioassay used in the study

|S|	Bioassay	No of Attributes	Screening type	No of Compounds
1	AID686978, AID686979	15	Confirmatory	848038+
2	AID686970, AID686971	14	Confirmatory	729702 +
4	AID492972, AID492953 & AID492956	4	Confirmatory	1006380+

3.2 Data Preparation

In this model author need to classify the active and inactive compounds in the Bioassay. Before moving into feeding data into to the neural network, it has to be prepared and for data preparation Author have used Anaconda interface, it is a free open-source platform distribution of the Python and R programming languages for scientific computing that aims to ease up the package management and deployment. Anaconda has a virtual environment manager, called Anaconda Navigator, so it eliminates the need to learn to install each library independently. Major perks of using Anaconda is the presence various kinds of packages, it consists of about 1,400 packages and much more Conda libraries that makes coding process much easier less time consuming.

In the process of data preparation, cleansing of the dataset is done. It involves the removal of the unwanted data, unimportant columns and correction of the missing values.

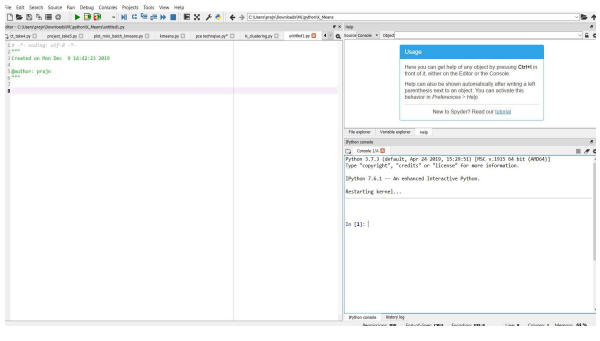

Fig.1. Spyder user interface

For the implementation of the algorithm I have used Spyder, a powerful scientific python integrated developmental environment. This platform allows the advanced combination of editing debugging and even testing of the data. It even allows the extension of its libraries using various plugins that a developer can fully exploit.

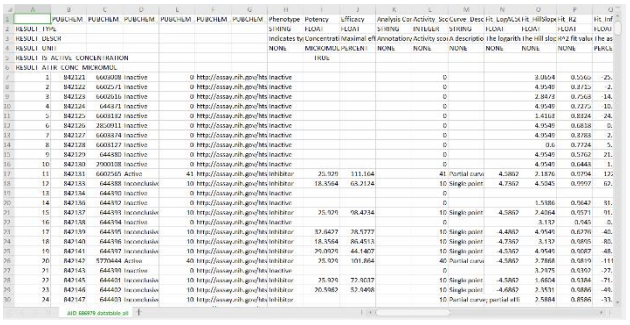

Fig.5. Original dataset with all the wanted and unwanted data columns.

Above figure shows the actual dataset. The entire dataset is loaded into Spyder. The dataset has many unreliable columns.

There are several columns and rows which are totally unnecessary for the classification. The data set at the very beginning has the few rows that gives information about the respective data sets. These rows are deleted can be deleted manually or by using the del function int the python. Author in the project deleted them manually. Column containing the missing values and there were columns which did not had any values or just the normal comments were also removed from the dataset even though these values could be replaced with zeros but might lead to inconsistency in the later prediction. The original data consisted of 47 columns and more than three hundred thousand rows and after processing it is reduced to 15/14 columns values are converted to integer. The image below shows the prepared dataset.

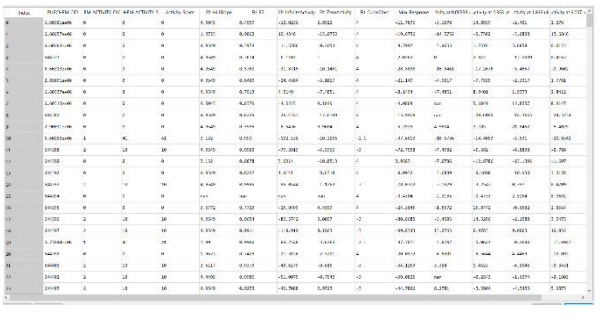

Fig.2. Dataset after pre-processing

Now the dataset has the pre-processed (Fig2) values which are essential for the prediction process. But still there are some columns that have null values, but they might not cause much variance in the classification. Data values are in floating point.

3.3 Data Processing

Once the data value is prepared now, they ready to be fed into the neural network. It includes all the essential features and also the required data format. At this stage begins the implementation of the neural network model to the dataset to test how good neural network can classify the bioassay compounds. Scikit-learn comes handy at this stage, a free software machine learning library for the Python programming language. It comes with useful features that are used in various classification, regression and clustering algorithms. In this project I have used Scikit-learn for coding that is good in interoperating with NumPy and SciPy which are Numerical of Scientific libraries of python. Author started off with creating the supervised neural network, where author considered fourteen different feats except for the activity_outcome as the independent feature based on which the analysis was done.

3.3.1 Neural Network

Neural network is a series of algorithm that executes to recognize the relationship between the set of underlying data. More like it mimics the work of human brain to understand from it's given input. It adapts to the changing inputs and outputs a best possible result. It bears a strong similarity to statistical methods like curve fitting and regression analysis.

Once a enough examples have been processed during training, it uses this as an answer key to understand the characteristic of the inputs while predicting to give the best output. The neural network used in this project consists of five hidden layers and it computes a total of 150 iterations. Author have worked with ReLu activation function and the Sigmoid neurons are used as the it gives a smother output than a common perceptron neuron. in this model learning rate is kept constant throughout the network. Stochastic gradient descent is applied for better weight optimization. Below is results that were obtained after application of the neural network for the bioassay datasets.

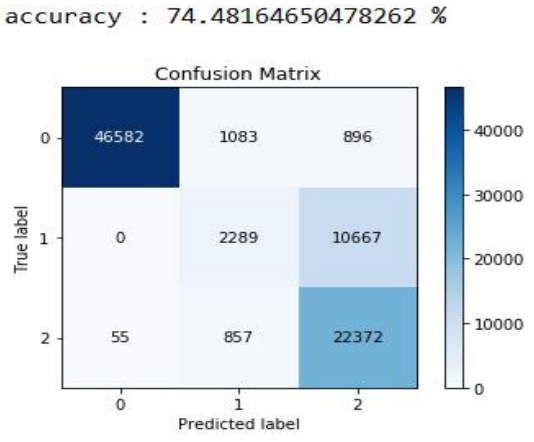

Fig.3. Accuracy and Confusion metrics Bioassay AID686978

Most classifier assume equal weighting of the classes in terms of number of instances and importance. Hence misclassifying one input has the weight as misclassifying another. This leads to a lot of false positives. In neural network loss function is used to calculate the loss or error in the prediction at each neuron. This loss function helps to keep the misclassification cost low or zero.

accuracy : -0.8593820466822111 %

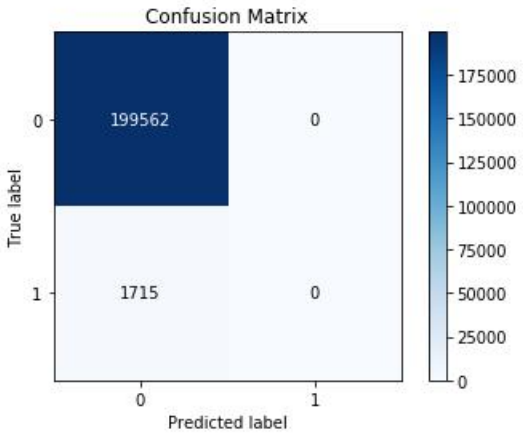

Fig.4. Accuracy and Confusion metrics Bioassay AID492953, AID492956, AID492972

In above confusion metrics Fig.4 there is a good number of data have been classified as inactive as shown in the dataset. But when it comes to classification of active compounds all the compounds are miss classified as inactive. This gave more than thousand false positives for the active compounds. This problem might have caused due to not fitting the training data into the neural network model. This huge range of misclassification lead to a negative accuracy.

The above confusion metrics Fig.5 is for the same set of data when training and testing data are fit into the neural model. This gave very good improvement in the prediction. The number of compounds that are misclassified are way less than before when the data was not fit into the prediction model. As the data set has the only two major features it does not require other boosting techniques like dimension reduction to improve the accuracy. Though there were zero inactive compounds that was misclassified as active. But this the above confusion metrics has sixty compounds that are falsely classified as active. The same goes with false positives of active compounds.

accuracy : 93.26094637425196 %

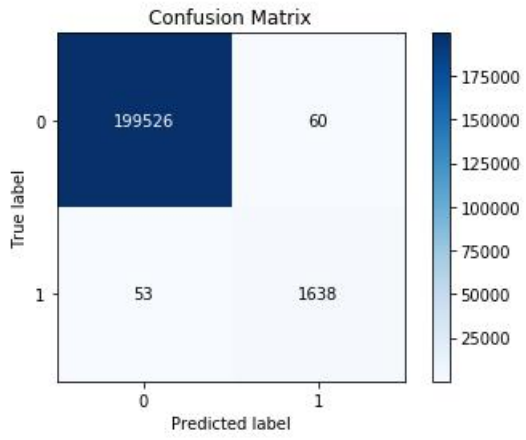

Fig.5. Accuracy and Confusion metrics Bioassay AID492953, AID492956, AID492972

In previous work they have combination of primary and confirmatory data sets in the experiment. This gave more chances of misclassified data that increased the number of false positives, were as in this paper author have used only the confirmatory data of a bioassay that consistently reduced the number of false positives and false negatives in the dataset.

accuracy : -58.96979411025758 %

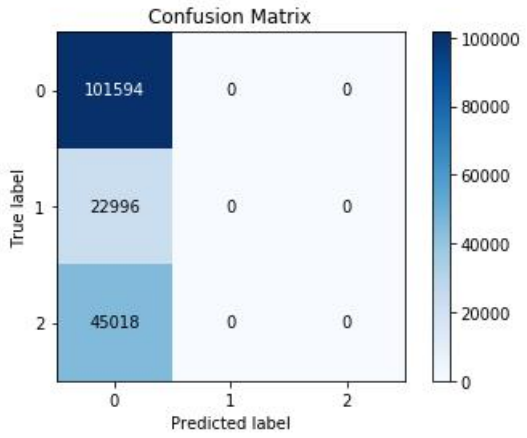

Fig.6. Accuracy and Confusion metrics Bioassay AID686978, AID686979

In the above image the confusion metrics of the bioassay AID_686978 and AID_686979 is given. As you can see the accuracy is very less at the first attempt of considering all the 14 features during the prediction. Almost all the values are falsely classified. There not even single value which is correctly classified in category 1 and 2 but incase of category 0 all the compounds are correctly classified.

Fig.7 shows the output of the neural network for bioassay AID686970 and AID686971. It is evident that the output is overfitting and the accuracy is not reliable. Hence passing this prediction model into dimension reduction seems to be a good idea. The number of hidden layers used in this model is five and outer layer uses activation function ReLu and hidden layer uses sigmoid activation function.

accuracy : 99.65411124728332 %

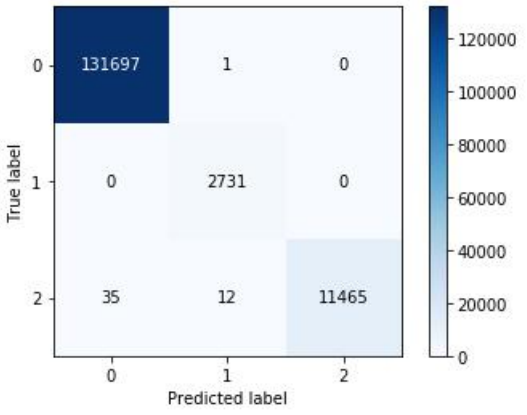

Fig.7. Accuracy and Confusion metrics Bioassay AID686970, AID686971

3.3.2 Dimension reduction

It is the common reduction technique for classification and other tasks. It is the process of obtaining the set of principle variable through minimizing the number of under consideration random variables. This approach can be divided into feature selection and feature extraction. Feature selection is the approach of selecting the set of input variables hence reducing the higher dimension features to lower to get better accuracy.

Two of the dimension reduction technique used in this experiment are PCA and LDA.

3.3.2.1 Principal Component analysis

It performs the dimension reduction by linearly mapping the data to lower dimension in such a way that, variance in the lower dimension used is maximized. Eigen vector of the matrix is calculated by constructing a covariance matrix of the data. In this way we can take the eigen value that corresponds to the largest eigen value to reconstruct the large portion of the original data set. Usually first few eigen vectors have the largest eigen value that can lead to a better accuracy in the prediction.

3.3.2.2 Linear Discriminant analysis

It is the most commonly used method in preprocessing of data in statistics pattern recognition and machine learning to find the linear combination of the features that can separate two or objects or events. This approach is almost similar to PCA, but the only difference is that it not only just recognizes the component axes that maximizes the variance but also the axes that maximize the separation between the multiple

classes. In this project dimension reduction was applied to 2 out of 3 sets of datasets below are the are the images of the confusion matrices after application of the dimension reduction

After application of dimension reduction there was a consistent improve in the accuracy. In the given confusion metrics Fig.8.shows the impact of the dimension reduction on bioassay classification. Confusion metrics shows a consistent reduction in the number of false positives in the output. Author have set the number of components used at a time of prediction to two. With number of iterations that neural network is being set to 200 and hidden layers being 6.

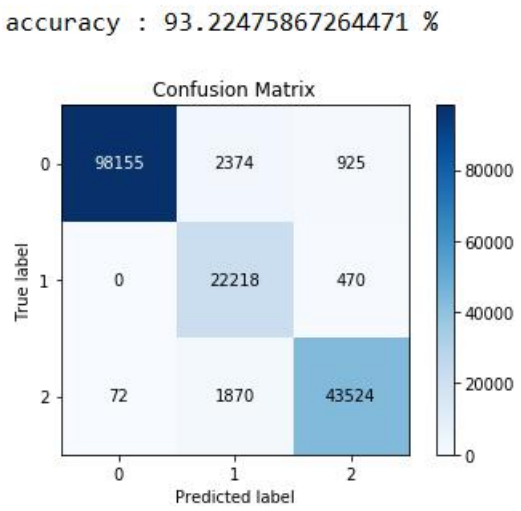

Fig.8. Accuracy and Confusion metrics Bioassay AID686978, AID686979

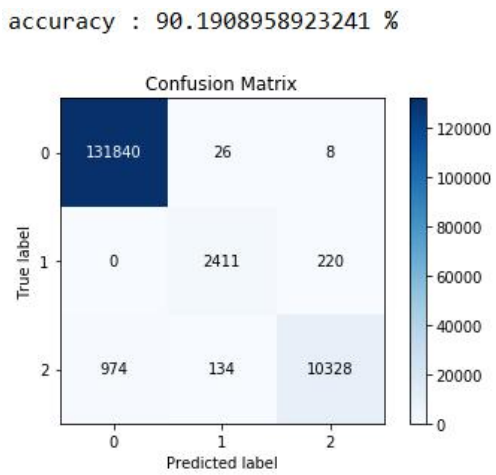

Fig.9. Accuracy and Confusion metrics Bioassay AID686970, AID686971

Fig.9 gives the visualization of the confusion metrics of bioassay AID686970 & AID686971 and there us good reduction of overfitting. The dimension reduction applied in this technique is uses three components as feature set. There is a good number of false positives and false negatives, but this can be improved with increasing the number of hidden layers and also number of iterations. Zero percent of the compound is classified into false positive inactive and the same goes for the compounds that are cytotoxic.

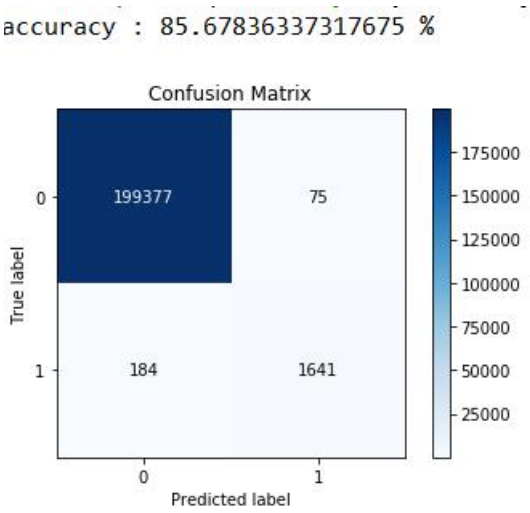

Fig.10. Accuracy and Confusion metrics Bioassay AID686972, AID492953, AID492952

Fig.10 shows the output of the neural network that works with dimension reduced data. The accuracy showed to be 85.67% but this accuracy is quite less keeping in mind that dimension reduction is used. As the number of features used in this dataset is only three application of dimension reduction to such dataset set did not seem to be a good idea.

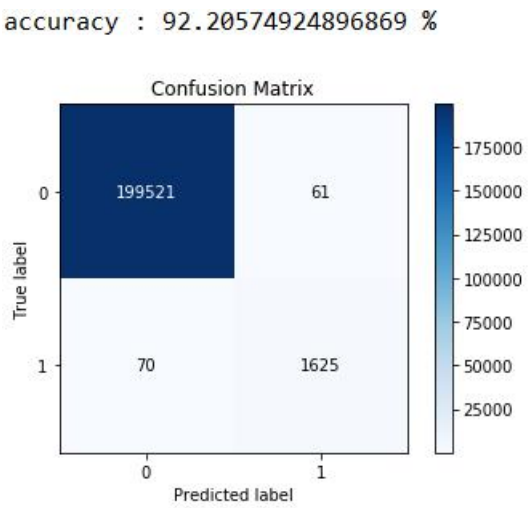

Fig.11. Accuracy and Confusion metrics Bioassay AID492972, AID492953, AID492956

Fig.11 is the output of the same set of the dataset without the application of the dimension reduction. Previously Linear Discriminant Analysis is applied. Neural network works better without the LDA in case of this dataset. Among Linear Discriminant Analysis and Principle Component Analysis, PCA seems to work way better than LDA. In all of the three set of datasets used LDA is leading to a overfit.

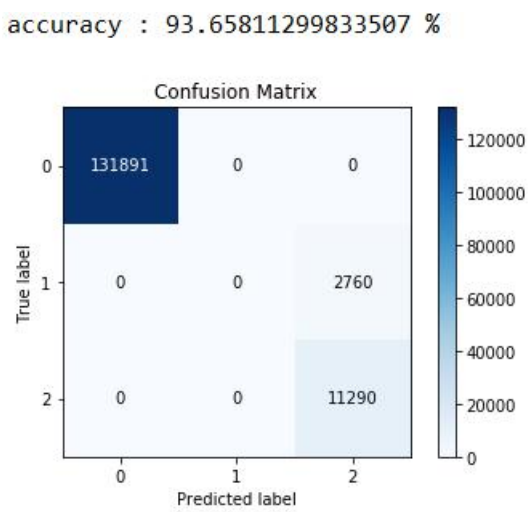

Fig.12. Accuracy and Confusion metrics Bioassay AID686970, AID686971

Fig.12 is the visualized confusion metrics for the bioassay AID686971, AID686970. Apart from other two bioassay dataset this dataset worked better with Linear Discriminant Analysis. When LDA was applied on this dataset it improved its classification accuracy to 93.65%. PCA has reduced the accuracy of model from 99% which was an overfitting to an accuracy of 90.19%. This result had a greater number of false negatives. From the Fig.12 you can see that active and inactive and inconclusive compounds are well classified. But the only problem is with classification of the active compounds.

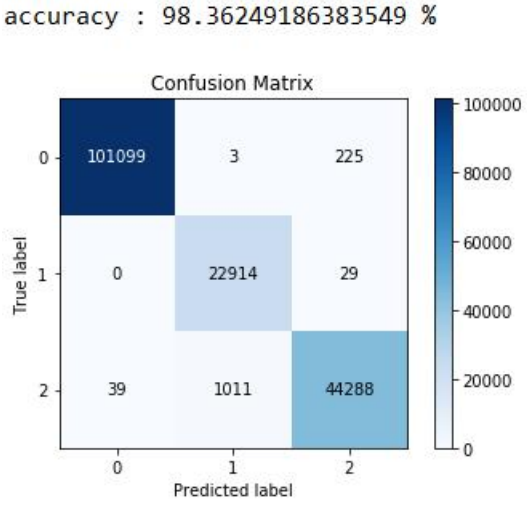

Fig.13. Accuracy and Confusion metrics Bioassay AID686978, AID686979.

With increase in the number of hidden layers to 6 and number of iterations to 200 there is a steep increase in the classification accuracy of the compounds as shown in the Fig.13. This configuration produced a drastic reduce in the number of misclassified compounds. The same effect was seen in the other set of the data as well. With increase in the number of hidden layers and the number of iterations, varying response of the different dataset ca be seen.

3 Result and Discussion

Table 2. Accuracy list for different sets of bioassay data

|S|	Bioassay	Accuracy Layer = 5	Accuracy Layer = 6	Accuracy Layer= 8	Accuracy Layer = 10
1	AID686978, AID686979	93.22%	98.36%	82.68%	79.97%
2	AID686970, AID686971	85.67%	89.74%	90.32%	89.45%
4	AID492972, AID492953 & AID492956	92.20%	93.94%	93.19%	93.87%

- For the first set of data used that is AID686978 & AID686979, neural network gave an accuracy of 93.22% as shown in the Table 2. Application of the dimension reduction gave a commendable improvement in the classification.
- For the second set of data used is AID686970 & AID686971. Neural network gave an accuracy of 93.65% with Linear Discriminant Analysis. PCA gave lesser accuracy of 90.19%. Even though there is a good accuracy, whole of Active category that is the category 1 in confusion metrics was misclassified.
- For last set of datasets, AID492972, AID492953 & AID492956, 92.20% accuracy was obtained from the neural network. There is a good number of compounds are classified correctly. As the size of feature set is 4, there was no necessity for the application of dimensional reduction. Confusion metrics shows that a small fraction of the compounds being misclassified. But this might be reduced using higher number of hidden layers in the neural network.
- Compared to the previous work this model does not suffer with lack of memory to handle the large set of data. Model is robust in terms of performance time and classification accuracy.
- This model shows a lower rate of misclassification compared to the previous work where it's method of handling the misclassification with reweighing the misclassification cost from the confusion metrics is not promising.
- With the increase in the number of hidden layers to 6 first set of datasets that is AID686979 & AID 686978, gave an accuracy of 98.36% along with Principal component analysis.
- But the same result was not seen in the second set of data (AID686971 & AID686970), where the Principal component gave a accuracy of 89.74% which is comparatively better than the previous performance of PCA. This time LDA gave an accuracy of 99% which seems to be overfitting. At this point author came to understand that LDA is more likely to overfit the classification.
- Last set of data (AID492972, AID492953 & AID492956), showed increase in the performance as the number of hidden layers increased. There is decrease in the rate of misclassification. Accuracy for this configuration was 93.90%. is possible that the model will perform even better with increase in the number of hidden layers and the iterations.
- The accuracy for the first set of data is highest with hidden layer number being 6 and as the number of hidden layers increased the model did not perform well. The same trend was seen in third set of datasets. But for second set for 8 number of hidden layers model gave an accuracy of 90.32% later at 10 number of hidden layers accuracy again reduced.

4 Conclusion

After the application of the neural network in virtual screening of the bioassay dataset for selection of compounds in high throughput screening. The model performed well in the given configurations. After examining the three different sets of datasets that includes seven various kinds of bioassay data the number of misclassifications was reduced in good number in comparison to the previously done work that includes the application of the Weka cost sensitive classifiers. The use of confirmatory screened data proved to be more efficient than primary screened ones that helped in reducing the percentage of missing values in dataset. In terms better performance neural network is a best option when bioassay data is large. But for smaller data Weka classifiers could be a best fit. But neural network is more efficient when it comes to weight optimization and misclassification reduction which was the main concern in Weka classifiers. With this for virtual screening of bioassay data neural network model is more recommendable.

Acknowledgements

I would like to thank Prof. Haiyan Hu for providing me the opportunity to work on this project and also National Database system for maintaining the data regarding different bioassays.

References

Schierz AC. Virtual screening of bioassay data. *J Cheminform*. 2009; 1:21. Published 2009 Dec 22. doi:10.1186/1758-2946-1-2.

Tufts Center for the Study of Drug Development, Tufts University, 192 South Street, Suite 550, Boston, MA 02111, USA.

Dealing with a data dilemma. *Nat Rev Drug Discov* **7,** 632–633 (2008). https://doi.org/10.1038/nrd2649

Lin JS, Ligomenides PA, Freedman MT, Mun SK. Application of artificial neural networks for reduction of false-positive detections in digital chest radiographs. *Proc Annu Symp Comput Appl Med Care*. 1993;434–438.

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Chem. Inf. Model. 2007, 47, 2, 264-278 Publication Date:January 9, 2007 https://doi.org/10.1021/ci600289v