Paper 266-25

Using Enterprise Miner™ to Explore and Exploit Drug Discovery Data

Michael S. Lajiness, Pharmacia & Upjohn, Kalamazoo, MI

ABSTRACT

One of the biggest challenges in modern pharmaceutical drug discovery is the effective management and exploitation of research data. How can one find those relatively small nuggets of knowledge buried in the great morass of data being generated by High Throughput Screening systems? One potential answer is by using Enterprise Miner™ software! This talk will describe how it can be used to find structural features of potential drug molecules that appear to effect interesting (and useful) biological activity. It will be shown how one can use Enterprise Miner™ even at a very basic level to uncover potentially useful relationships that can be exploited to more rapidly find new therapeutic candidates. Special issues with respect to how to mathematically describe chemical structures will also be addressed. This talk is geared towards SAS users at the beginning to intermediate skill level.

INTRODUCTION

Modern pharmaceutical drug discovery typically utilizes screening of many thousands of compounds in a controlled biological assay to determine potentially efficacious treatments. The most effective compounds, called “hits”, then may be tested again and/or subjected to secondary (follow-up) testing. The best “leads” may result in a chemical synthesis program to optimize the molecular structure with respect to the biological activity of interest.

It was the goal of the present work to become familiar with Enterprise Miner™ (EM) and determine how it may be used to make the discovery process more efficient and effective. Based on the results of this study, EM is likely to become an integral tool in pharmaceutical drug discovery at PNU.

To demonstrate the ability of EM to effectively explore and exploit drug discovery data it was decided to analyze a set of 1,648 monoamine oxidase results graciously provided by Yvonne Martin of Abbott Laboratories (Abbott). Specifically, it will be demonstrated how to

- Specify the input dataset
- Partition the file into training, validation and test sets
- Generate models based on logistic regression, decision trees, and neural nets
- Assess the performance of the various models
- Use the “best” model to predict the performance of new data and to evaluate it
- Add custom nodes to produce desired output
- Call external visualization software (SPOTFIRE) to help understand the results.

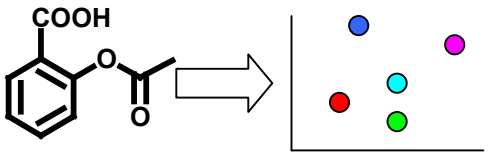
There are a lot of details that will not be delved into in this paper given the time and space limitations. We will instead focus on how one can easily use EM at a very basic level to generate some useful results. Please be aware that some of the more trivial details are glossed over so we can focus on more important aspects of using EM.

Background

One of the most significant choices one has to make to study the relationship between chemical structure and biological activity is how to represent the chemical structure. There are very many ways to do this and it is beyond the scope of a

SUGI paper to get into these details. One can think of the problem in general terms using the analogy of the blind men and the elephant. One man describes the elephant as having 4 large tree-like appendages. Another describes a long flexible snake (the snout!) with an opening at the end. The third describes the elephant as being huge or something like that. These descriptions are all right and all wrong at the same time.

So, how does one describe chemical structure information so that one can use mathematical / statistical tools to analyze it? Most often, one maps chemical structure into Euclidean (coordinate-based) space.



One can use fragment bits, topological indices, or more recently BCUTs. For several reasons it was decided to utilize BCUT representations for this paper. These are generated by DiverseSolutions software which is available commercially (DVS).

BCUTs

It has now been clearly and repeatedly demonstrated (by numerous groups in pharmaceutical and agrochemical industry) that "BCUT values" (vide infra) are useful descriptors of chemical structures

In 1989, Burden suggested that a "molecular ID-number" could be defined in terms of the two lowest eigenvalues of a matrix representing the hydrogen-suppressed connection table of the molecule. Subsequently, Rusinko and Lipkus and then later on Pearlman and Smith added significant extensions which resulted in what is now referred to as the BCUT approach (Pearlman, et al. 1998). The basic approach attempts to describe molecules with respect to the way they might interact with a bioreceptor. Some of these descriptors relate to atomic charges, atomic polarizabilities, and atomic H-bond-abilities of the compound. This results in 3 types of BCUT descriptors.

DiverseSolutions software written by Bob Pearlman at the University of Texas at Austin was used to generate three types of BCUTs: 2-dimensional; 3-dimensional; and 2-d-topologic. The 2-d BCUTs are based on the 2-dimensional representation of the chemical structure and the corresponding adjacency matrix. The 3-d BCUTs are defined in terms of the 3-dimensional structure of the molecules as determined by the CONCORD program also written by Pearlman’s group. The 2d-topologic BCUTs are based on the 2-d topology of the molecule.

The Problem

Let’s say that we have just run a biological assay, which has generated the Monoamine Oxidase (MAO) data, on a set of

structurally diverse compounds that is a representative subset of a larger collection. This is often done in the pharmaceutical industry to conserve resources. Now we want to derive a model that will identify the most active compounds contained in the remaining collection that have not been screened. Obviously, we also need to verify that whatever model is chosen produces reasonable results in terms of correct predictions.

Data Set Variables

In the input dataset there are 16 2-d BCUTS (BCT2D1-BCT2D16), 18 2-d-topologic BCUTs (BCT2DT1-BCT2DT18); and 29 3-d BCUTs (BCT3D1-BCT3D29). The biological response variable for the MAO dataset is Log Score (SCORE) which is measured on a range of 0 to 3, 3 being most active. In addition to SCORE, 2 other activity fields are defined to facilitate additional modeling where one categorizes activity as either Active (A) or Inactive(I). The field ACT was defined to be I for scores of 0, and A for scores>0. ACT2 was defined to be I for scores less than 2 and A for scores>=2.

USING ENTERPRISE MINER™

In the next several sections we will go through the various steps to create, modify and run an EM project to analyze the MAO data. Due to time and space limitations not all corresponding screen shots or details will be mentioned but the main points will be hit so that the reader can understand the application. Please note that the following sequence of operations is very similar to the example process flow diagram used in the Version 3, Getting Started with Enterprise Miner™ Software publication.

The basic strategy we generally follow when trying out new software and techniques is to try and develop the simplest application possible and then fine tune. What that means in this example is that we will build a very basic EM diagram to analyze the MAO data. We will not transform, filter, or do any variable selection to reduce dimensionality. We will use the default settings for most nodes. The results that come out of these efforts then form a “baseline” that we can attempt to improve by various methods. We have found this to be an effective way to get “into” new software and to make a positive impact quickly.

In general, one can run EM in basically two ways. You can build the diagram, one node at a time, and then run each node by right clicking on the icon and selecting RUN. Or one can wait until a “later” node and then select RUN there. Whenever RUN is selected all predecessor nodes that have not been run are executed.

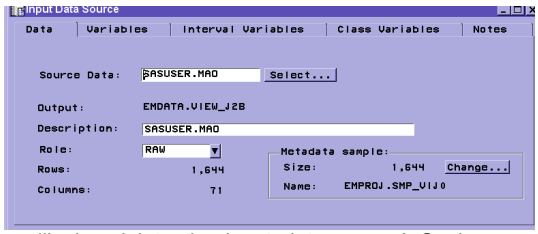
Creating the EM Project

One needs to create a project by selecting File>New>Project in the EM screen. I created the project MAO and chose to put it in the default location. Choosing the default location makes it easy to find the project later on. Once you click on the CREATE button, EM will automatically create an untitled diagram which you can then start customizing. At this point you simply type in the name of the diagram you are going to create which can be thought of as a particular set of analyses. A project can have many different diagrams.

Basically the various steps in the diagram or datamining project we are defining are based on the SEMMA paradigm: Sample; Explore; Modify; Model; and Assess.

Defining the Input Dataset

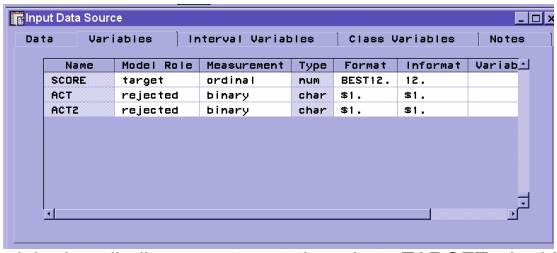
One of the neatest things about EM is how one can use visual programming icons to do almost all of the construction of the project. To define a dataset, simply click and drag the INPUT DATA SOURCE icon from the tools box to the project area. Double click on the icon to define the input dataset. It is possible to store the EM input datasets virtually anywhere. However, it is convenient to store them in a directory that will automatically be available. Thus, I chose to place the MAO dataset created above in the EMDATA subfolder for the project. Please note, however, that if you delete the project



you will also delete the input data source! So be careful! Once the dataset is in the proper directory one can click on SELECT and then one can click on the MAO dataset.

Setting the TARGET Variable

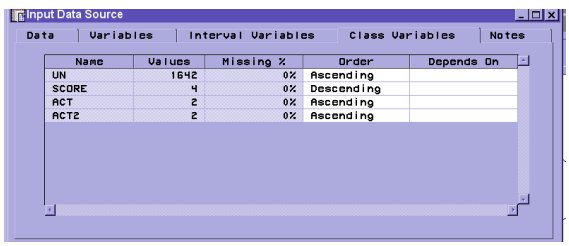
Modeling requires the existence of at least one target or dependent variable. To select the appropriate target in the MAO dataset one needs to click on the VARIABLES tab and scroll down to the SCORE variable. Right clicking in the



model role cell allows one to set the role to TARGET. In this application I chose to set the roles for ACT and ACT2 to “REJECTED” so that they are not used in any modeling.

Setting the Event Level

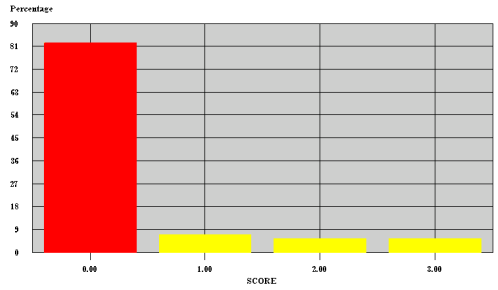
One needs to make sure that EM recognizes what event to model; thus, one needs to properly define the event level for the target variable. Since the SCORE variable is an ordinal variable, and the highest score, 3, is the best score, we need to define ORDER as DESCENDING. If, on the other hand, the best score was zero we would have to set the order to ASCENDING. Care needs to be taken when setting the event level for character/binary variables so that the order chosen reflects the proper sort sequence. For example, if we used the variable ACT as the target and wanted to generate models to best predict active compounds (ACT=A) then one would use an Ascending order for ACT.



One can also define a target profile for Score. Using this, one can define decision matrices and prior probabilities. More information on target profiles can be found in the online guide.

Viewing the Distribution of Score & other Variables

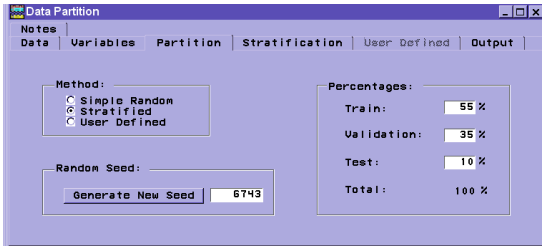
It is always a good idea to look at the distribution of your target variable as well as the other variables. You can get a quick and easy plot by right clicking on the selected variable under the VARIABLES tab and selecting VIEW. An example of the distribution of SCORE is shown below.



Variables that are heavily skewed or have large kurtosis values as indicated in the data displayed under the INTERVAL VARIABLES tab may be transformed or filtered to remove outliers. There are both a TRANSFORM VARIABLES NODE and a FILTER OUTLIERS node. While many of the analysis methods utilized in EM are fairly robust with respect to violations in the underlying assumptions of normality, homogeneous variance, and additivity, it is still a good idea to check the distributions of the variables and to transform them if need be. Since the BCUTs are already standardized in some way, it was decided not to perform any transformations or filtering.

Partitioning the Dataset

In developing models it is usual practice to split the dataset into 3 parts; a set to develop or train the model; a set to validate the model; and a set to independently test the model. The validation step is necessary to prevent overfitting. To include a partitioning step in our diagram we simply click and drag a connection from the INPUT DATA SOURCE node to the DATA PARTITION node. In the MAO dataset, we have chosen to partition the dataset as indicated in the figure below. Note that we have reserved only 10% of the data in the TEST set to get an independent test of the final model.



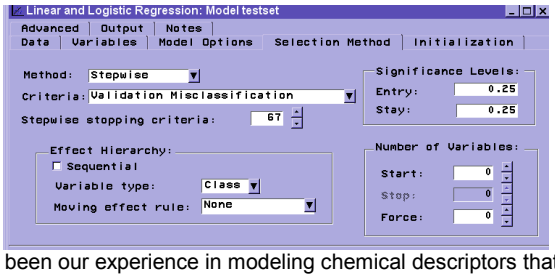
Also note that in addition to the percentages, the STRATIFIED method is chosen using the SCORE variable so that the proportion of compounds with the various scores are the same across the three sets.

Modeling

There are three main models one can access via the MODEL section of EM: Regression, Neural Networks, and Decision Trees. It is possible to add others as well.

Regression Model

After dragging and connecting the Regression node to the Data Partition node we, double click on the icon to open it. The USE status of the variables is exactly as set earlier. For the MAO dataset we modified the SELECTION METHOD as shown below and requested a STEPWISE regression.

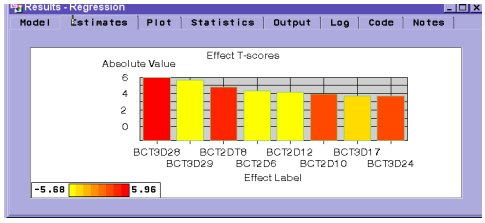


It's been our experience in modeling chemical descriptors that one obtains better results with a significance level for keeping variables in the model at about .25. Please note that we have also set the criteria for evolving the model to "Validation misclassification", so that we will find the best model that minimizes the misclassification rate on the validation dataset.

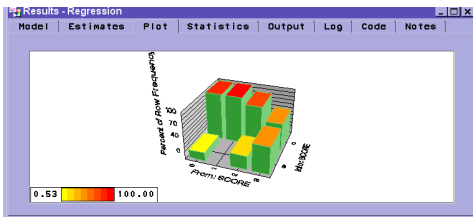
One can save this current model using the "File > Save Model As" pull-downs under a given name. It is possible to create several different models and to recall them later. Otherwise, the current model will be replaced the next time you modify the settings and run the diagram.

At this point it is worthwhile to mention the NOTES tab. It is useful to enter information into the NOTES tab for each node whenever you select something other than the default. One can place comments in there as to why a particular choice was made with reference to other published material. This becomes quite useful when one utilizes the REPORTER node as will be described later. It would be useful to be able to user hyperlinks here but that does not appear to work since the notes file is perceived to be pure text.

Once this regression mode is run you can examine the EFFECT T-scores, which give a plot comparing several of the most important effects. This particular plot shows that the 28th 3D BCUT (BCT3D28), which is related to low values in the polarization of the compound, is the most important variable in the model. This polarization BCUT is essentially a measure of the "greasiness" or lipid-like nature of the compound. One might infer that by decreasing the lipophilicitiy of active compounds one could enhance activity.



One can also examine the "From & Into" plot that shows you how well the model worked with respect to the training set. "From" corresponds to the observed score and "Into" refers to



the predicted score.

One can observe the Training set results, which is at the end of the report under the OUTPUT tab. If one switches to the output tab clicks on the slider bar and then presses the END key you can immediately go to the end of the file. It should be noted that only the training set results table appears under the OUTPUT tab.

Logistic Regr. Training set

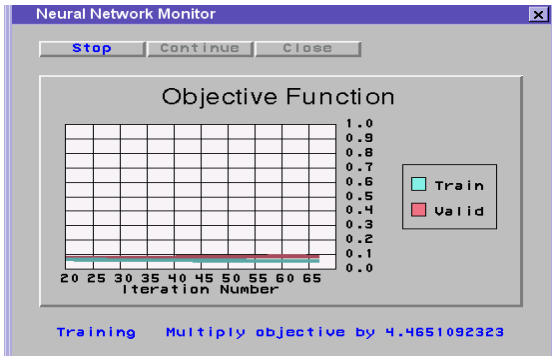
From \ Into	0	1	2	3
0	752	0	0	4
1	63	0	0	0
2	33	0	0	3
3	24	0	0	24

Neural Network Model

In our experience at PNU Neural Networks (NNs) can be applied to many problems in drug discovery and they are a worthwhile addition to the EM toolkit. That said, my own personal bias is that neural networks are extremely complex and there are many different architectures that one can build and many ways to influence convergence and the predictive results. They offer a nice "black box" approach if you are lucky enough to get something predictive that appears to work but the resultant parameters and weights will probably not give you much insight into the problem you are studying. In terms of drug discovery, a good neural network that effectively models activity as a function of structure will not, in general, give a chemist very much insight into how one should one design the next compound to make a more potent inhibitor.

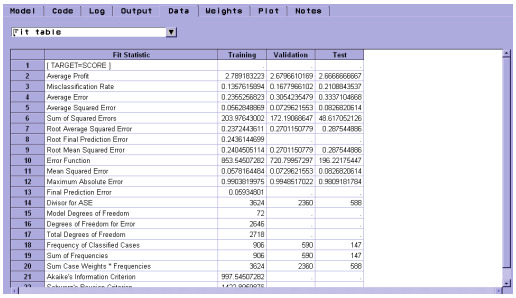
Anyway, after dragging and connecting the Neural Network Node to the Data Partition node one can open it. Note that at this point both the Regression and NN node are connected to the partition node. Under the GENERAL tab we can set the Model Selection criteria to MISCLASSIFICATION Rate just as we did for regression. In the present example, we are just accepting the basic NN model. Please note that one can quite often significantly improve NN performance by changing the architecture and other settings available under the ADVANCED tab.

However, it was desired in this case to illustrate how one can use the most basic features of EM to get useful results. Now you can train the model by selecting TRAIN MODEL under the TOOLS menu, or back out and then right click and RUN. You then get to see the progress of the model as shown below.



Once the model is completed you can examine the results.

Unfortunately, one can only see the output as shown below after successful training. No table of prediction results for training, validation or test sets comes out directly. One can obtain these tables, and a way to do this will be discussed later.



Decision Tree Model

After dragging and connecting the TREE node to the Partition node it is ready to run! The assessment criterion is already set to "Proportion correctly classified", which refers to the validation dataset. After running TREE one can examine the results under the summary tab. The summary table lists the classification results for both the Training and Validation sets but not the TEST set. These results are extracted below.

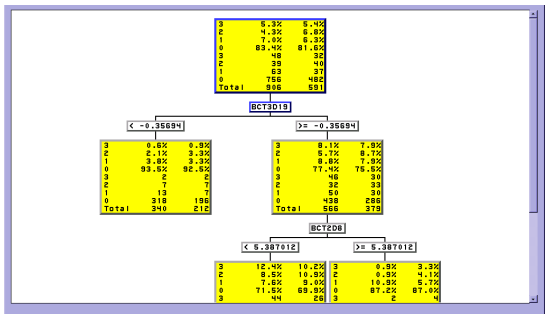
Decision Tree Training Set

From \ Into	0	1	2	3
0	754	0	1	1
1	54	9	0	0
2	31	0	5	3
3	21	0	0	27

Decision Tree Validation Set

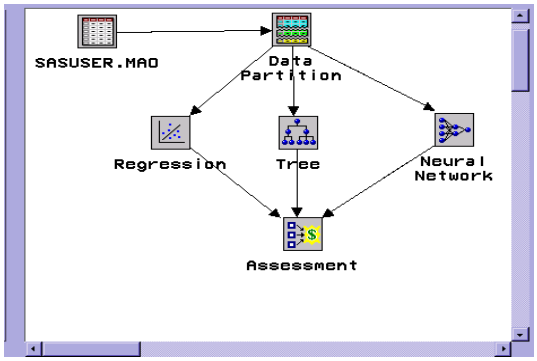
From \ Into	0	1	2	3
0	472	4	1	5
1	33	0	2	2
2	32	0	3	5
3	11	0	2	19

While viewing TREE results one can select VIEW > TREE to see the actual tree diagram as shown below. In this diagram we can see that the 19th 3-D BCUT is the most important variable, followed by 2D-BCUT number 8. As before, since each BCUT corresponds to a particular chemical feature known to be important in binding to a receptor, this knowledge should give a chemist ideas as to how to make a better compound.

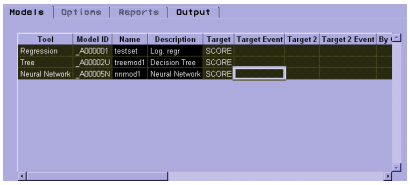


Assessing the Models

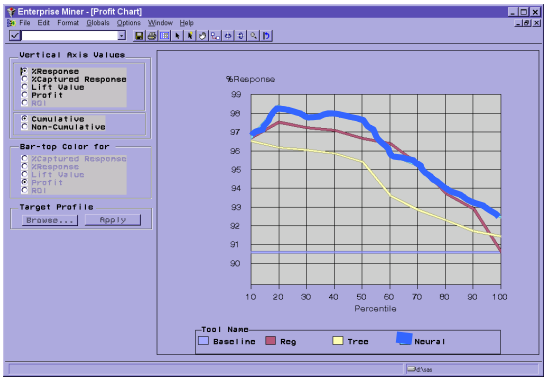
To assess the various models one can drag and connect an ASSESSMENT node to each of the analysis nodes. The partially completed tree now looks like the following diagram.



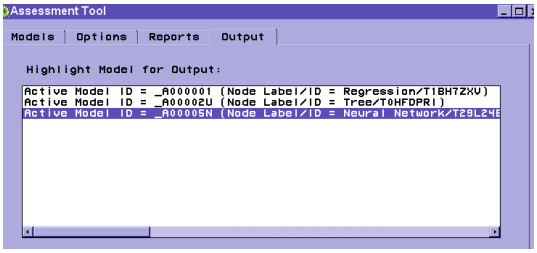
After opening the ASSESSMENT node you need to click, drag and select all three models under the MODELS tab.



Then, to perform the assessment, use the TOOLS pull down menu, select LIFT CHART, and select %Response, to obtain the chart below. From this chart it appears that the Neural Network model captured more of the "best" compounds more quickly than the other methods did.



Now, since we are interested in scoring we need to select the best model here by closing the lift chart window, selecting the OUTPUT tab, and then selecting the neural network model for "output".

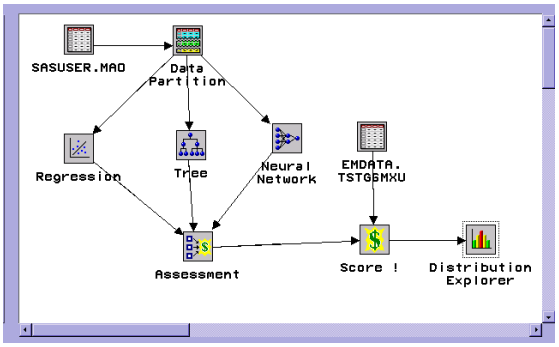


Defining and Scoring the Test data set

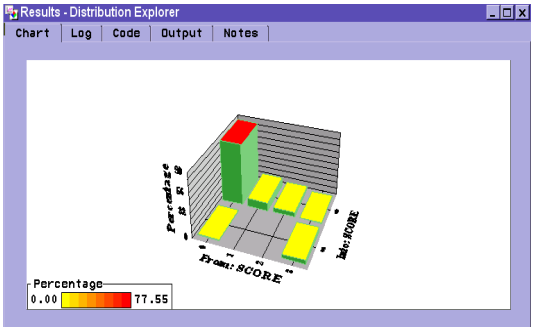
The Score node generates and manages scoring code from the models that were run. First, though, we have to tell EM what dataset to score. This is done through another INPUT DATA node. Note this node is not connected to anything yet, but WILL be connected to the SCORE node. We can open the node and drill down (or type in) the name of the dataset containing the data to be scored. The data we want scored for the MAO dataset, which contained 10% of the original dataset is named EMDATA.TSTG6MXU. You can identify the name of this file a number of ways, but one way is to OPEN the DATA PARTITION node and to look for the PROC CONTENTS section associated with the creation of the testset. One also needs to set the model role of the SCORE variable to PREDICT. Now, drag and connect a SCORE node to the ASSESSMENT node and the INPUT DATA node just created. Then under the SETTINGS tab of the SCORE node set the action to "Apply Training Data Score code….". Since we explicitly selected the neural network model to be output the corresponding scoring code is accessed.

View Predictions of "Test" Data

In the previous few nodes we input the test dataset and then scored it. We can now use DISTRIBUTION EXPLORER (DE) to see the results. We click and drag the DE node over and connect it to the SCORE node so our completed diagram (so far) looks like the following.



After we open the DE we can then select the variables (under the variables tab) that we wish to plot. We can then right click on the desired variables, F_score (original score) and I_score (predicted score), and set the axes to X or Y. We run the DE node and see the picture below whose corresponding tabular results are found under the output tab.



Even though the NN seemed to be the best model on the basis of the analysis discussed earlier, it is evident from the results below that the NN model leaves much to be desired. How would the other models have performed?

Neural Net 10:03 Mo
test results

TABLE OF F_SCORE BY I_SCORE

F_SCORE(From: SCORE) I_SCORE(Into: SCORE)

Frequency	0	2	3	Total
0	110	2	4	116
1	15	0	0	15
2	8	0	0	8
3	2	0	6	8
Total	135	2	10	147

What to do next?

It is evident that one needs to construct an EM diagram that can easily produce summarized results on all the models. Thus, a SAS CODE node was added to the diagram for each of the modeling nodes, as exemplified by the code below. In this node under the MACROS tab and the DATA SETs /ALL folder, the macro names for the training, validation and test sets can be found (&_mac_1, &_mac_2 and &_mac_3).

```
title1 'Neural Net';
proc freq data=&_mac_1;
tables f_score*i_score /
nopercent nocol norow nocum;
title2 'training results';
proc freq data=&_mac_2;
tables f_score*i_score /
nopercent nocol norow nocum;
title2 'validation results';
proc freq data=&_mac_3;
tables f_score*i_score /
nopercent nocol norow nocum;
title2 'test results';run;
```

If one runs the above code for each of the models, the PROC FREQ will tabulate the results of using each model to score the TEST dataset and produce the following tables. (Note: the test set results for NN were already displayed above)

tree 14:49 Fri
test results

TABLE OF F_SCORE BY I_SCORE

F_SCORE(From: SCORE) I_SCORE(Into: SCORE)

Frequency	0	1	3	Total
0	115	1	0	116
1	14	0	1	15
2	7	0	1	8
3	1	0	7	8
Total	137	1	9	147

logistic regression 12:3
test results

TABLE OF F_SCORE BY I_SCORE

F_SCORE(From: SCORE) I_SCORE(Into: SCORE)

Frequency	0	3	Total
0	113	3	116
1	15	0	15
2	8	0	8
3	4	4	8
Total	140	7	147

As one can see from the numbers, the NN did better than the logistic regression but not quite as well as the decision tree. The tree node identified 7 out of 8 of the most active compounds (score=3) and 8 out of 16 moderately active or above compounds (score=2,3). ***Thus, in our diagram we re-opened the ASSESSMENT node and changed the output model from NN to TREE.*** This makes the Tree-based model the one to be used for subsequent predictions.

Combining Models

Even though we have determined that the TREE is the best of the individual models, it is possible to use the ENSEMBLE node to come out with a consensus model that can be used to generate predictions. By dragging and connecting the EMSEMBLE node to the 3 model nodes and then to the ASSESSMENT node, one basically creates a new model by averaging the posterior probabilities (for class targets) or the predicted values (for interval targets) from multiple models. The new model is then used to score new data. By running the ENSEMBLE node in the present example we can obtain the following results for the test dataset.

TABLE OF F_SCORE BY I_SCORE

F_SCORE(From: SCORE) I_SCORE(Into: SCORE)

Frequency / Percent	0	3	Total
0	115	1	116
	78.23	0.68	78.91
1	14	1	15
	9.52	0.68	10.20
2	8	0	8
	5.44	0.00	5.44
3	1	7	8
	0.68	4.76	5.44
Total	138	9	147
	93.88	6.12	100.00

Note that these results are a bit better than the NN-based predictions but a bit worse than the TREE-based ones. Thus, one should examine the performance of the individual models when deciding whether to use an ENSEMBLE node. So, in the current diagram the ENSEMBLE node was dropped since results obtained with the TREE were superior.

Generating a list of the Best (Predicted) Compounds

Let's say that one wants to generate a list of compounds that we have the most confidence will have the desired biological activity, which in this case is monoamine oxidase activity. This is done by adding a new SAS CODE node and then defining what the cutoff is to be used to define confidence. So, a SAS CODE node was connected to the distribution explorer node. Now, one needs to define what compounds are to be output. That is, which compounds are likely to be the most biologically active. The way this is determined in the present example is by using DISTRIBUTION EXPLORER to look at the distribution of the predicted score=3 (the variable P_SCORE3). Recall that a score of 3 indicated the most potent inhibitor. On the basis of this examination it was decided to use a cutoff of P_SCORE3>. .17 to define the predicted "best" compounds. As before, one can look under the MACROS tab and then the DATASETS / ALL subtab to find the name of the SCORE dataset, &_mac_4, that can be referenced in the code. Now, under the PROGRAMA tab of the new SAS CODE node, we can type in the following program.

```
data goodones; set &_mac_4;
if p_score3>.17;
file 'c:\cousin\pscore.lst' put @1 un;
title1 'results for best compounds';
proc print data=goodones;
var un p_score3 P_score2 P_score1
    score f_score i_score;run;
```

Running this node generates a display of the compound id's, predicted and observed scores for this set, along with a file, pscore.lst, that can be used in our CHEMINFORMATICS system to examine structure and inventory. It should be noted that nine compounds match the p_score3 >0.17 filter and all of these (100%) are active. Also, one should note that there were 8, 8, and 15 compounds in the score = 3, 2, and 1 categories respectively. So it seems that EM is pretty good at finding very active compounds but not so good at picking up the moderately active or less compounds. This could still be very useful if one wanted to examine commercially available databases in a search for active inhibitors and then purchase those compounds that meet the p_score>0.17 criteria. One could raise or lower the threshold for p_score3 depending on how tolerant of inactive compounds we wanted to be.

Predictions on best predicted cmpds 10:0

OBS	UN	P_SCORE3	F_SCORE	I_SCORE
1	15220	0.7500	3	3
2	16992	0.7500	2	3
3	17380	0.7500	3	3
4	18720	0.7500	3	3
5	18748	0.7500	3	3
6	19108	0.9474	1	3
7	20536	0.9474	3	3
8	20542	0.9474	3	3
9	20920	0.9474	3	3

Predictions on best predicted cmpds 10:0

TABLE OF F_SCORE BY I_SCORE

F_SCORE(From: SCORE)
I_SCORE(Into: SCORE)

Frequency | 3	Total
1 | 1	1
2 | 1	1
3 | 7	7
Total 9	9

Further Enhancements to the diagram

One can add a REPORTER node that can generate an HTML report summarizing the results of the modeling. This appears to be a useful starting point for quickly preparing documentation for the analysis. It includes the final diagram, input datasets and variables, the partitioning that was used, as well as summarization of the analyses performed.

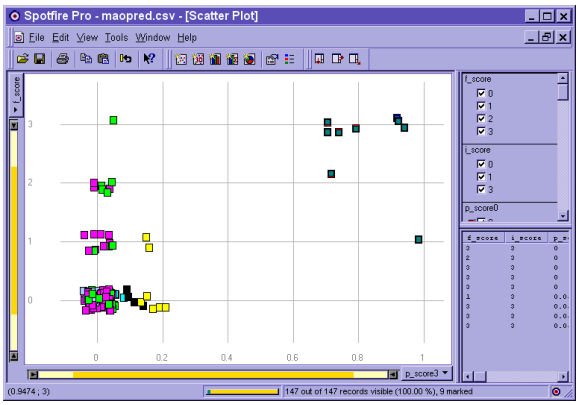
In addition, SPOTFIRE, visualization software that is very popular at PNU (www.spotfire.com), is often used to visually explore large amounts of data. Thus, we wanted to be able to export our EM results to SPOTFIRE. It will be illustrated how one can set up the input files necessary to run SPOTFIRE and then to execute the software from EM itself. It was decided to create 2 SAS CODE nodes. The first creates a .csv file containing the data to be exported, and the second invokes SPOTFIRE. The corresponding program code in the first node is shown below.

```
data _null_; set &_mac_4;
file 'c:\data\maopred.csv';
if _N_=1 then
put  'f_score'  ',' 'i_score'  ','
     'p_score0' ',' 'p_score1' ','
     'p_score2' ',' 'p_score3' ','
     'r_score0' ',' 'r_score1' ','
     'r_score2' ',' 'r_score3' ','
     'd_score'  ',' 'ep_score' ','
     'cp_score' ',' 'd_score'  ','
     'un' ;
put f_score  ','  i_score  ','
    p_score0 ','  p_score1 ','
    p_score2 ','  p_score3 ','
    r_score0 ','  r_score1 ','
    r_score2 ','  r_score3 ','
    d_score  ','  ep_score ','
    cp_score ','  d_score  ','
    un  ; run;
```

The second SAS CODE program that runs SPOTFIRE follows. Note that the RUN command issues a "windows" command, or in this case, run a windows program, i.e. SPOTFIRE.

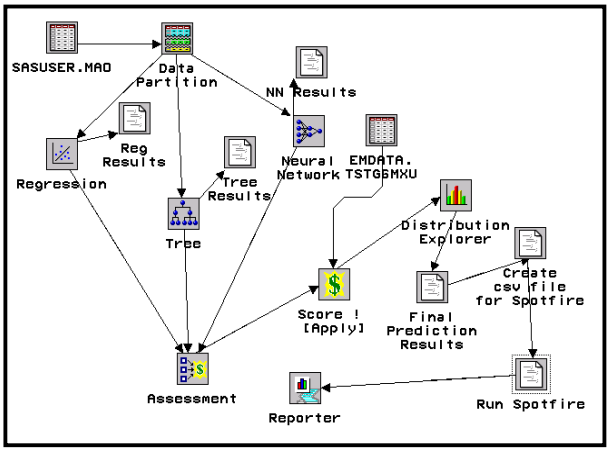
```
data _null_;
x d:\spotfire\program\spotfire.exe
  c:\data\maopred.csv;
run;
```

The results of these two SAS CODE steps are the creation of the .csv file, the import of the dataset into SPOTFIRE, and the automatic visualization of the appropriate data in SPOTFIRE. Running these nodes generates the corresponding SPOTFIRE plot.



Final Model

The final diagram that incorporates all the items we've discussed in this paper is shown below. Please remember that this is a very basic diagram and that one can generate far more detailed AND APPROPRIATE diagrams using more of the advanced features of the software.



Final Comments

When we first created the input dataset we included 2 additional variables that were excluded (rejected) from the analysis. These were the ACT and ACT2 variables, which were created to see if any improvement in prediction could be obtained by creating a binary activity flag. To run the diagram above on the variable ACT2, for example, one needs to change the model role in the first INPUT DATA SOURCE nodes to reflect the new target variable while setting the SCORE variable to REJECTED. Also in the SAS code node where the final prediction results are generated, one needs to change the variable names to the new one(s) that will be generated by analysis of ACT2.

It was found in the current example that no increase in predictivity was obtained by modeling ACT2 versus SCORE.

It should be reemphasized here that the results obtained with the MAO data through this application of Enterprise Miner™ could be dramatically improved. Use of the more advanced features available in the MODEL nodes or adding USER-DEFINED models to utilize other SAS procedures like DISCRIM and the new PLS (Partial Least Squares) would almost certainly improve our results.

In addition to user-defined models, one could also add other useful nodes such as a JMP® node and then execute JMP® in much the same way as we invoked Spotfire. The possibilities for enhancing and improving our application are almost unlimited.

For advanced users, the DM Tool provides an SCL programming interface for creating new tools for the SAS/Enterprise Miner environment. This is available in the Business Solutions category on the SAS.COM web site located under Software Demo's and Downloads. These tools may be used to provide additional customized functionality to end-users. Documentation and source code is provided there as well. The DM Tool class is included in the standard installation of SAS/Enterprise Miner v.3

CONCLUSION

Enterprise Miner™ appears to have great potential to make a significant impact in drug discovery. It is the opinion of the author that one can derive great benefits from using EM even at a very basic level. The many capabilities, options and intricacies involved with the product do make for a significant learning curve however. The visual programming paradigm is very cool and saves one from lots of coding. However, it seems evident that one cannot use version 3 EM without at least a little bit of coding in SAS. Thus, the software needs to evolve a bit more before EM can be used by those essentially unfamiliar with SAS code and SAS datasets.

The bottom line with respect to using Enterprise Miner™ on drug discovery data is a very positive one. Those involved with trying to get the most of drug discovery data should take a serious look at Enterprise Miner™.

REFERENCES

Abbott Laboratories. Upon signing an "agreement for confidential disclosure" Abbott Laboratories will make an electronic copy of this dataset available; contact Daniel W. Norbeck or Yvonne Martin, 100 Abbott Park Road, Abbott Park, IL 60064-3500.

DVS. DiverseSolutions™ is a commercially available software package offered by Tripos providing solutions to abroad range of diversity-related problems. Tripos is a leading provider of discovery research. Please see http://www.tripos.com/software/dvs.html for more information.

Pearlman, R.S., Smith, K.M. (1998). Novel Software Tools for Chemical Diversity. Perspectives in Drug Discovery & Design, 1998, 9, 339-353.

ACKNOWLEDGMENTS

I'd like to acknowledge the help of Veerabahu-Cons Shanmugasundaram who computed the BCUT values for the MAO dataset.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Mic Lajiness
Pharmacia & Upjohn
301 Henrietta St
Kalamazoo, Michigan 49008
616-833-1794(w):
616-833-9183(fax)
616-624-6690 (Mary's Bar & Grill)
michael.s.lajiness@am.pnu.com