Exploratory data analysis in the context of data mining and resampling.

Análisis de Datos Exploratorio en el contexto de extracción de datos y remuestreo.

Chong Ho Yu
Arizona State University

ABSTRACT

Today there are quite a few widespread misconceptions of exploratory data analysis (EDA). One of these misperceptions is that EDA is said to be opposed to statistical modeling. Actually, the essence of EDA is not about putting aside all modeling and preconceptions; rather, researchers are urged not to start the analysis with a *strong* preconception only, and thus modeling is still legitimate in EDA. In addition, the nature of EDA has been changing due to the emergence of new methods and convergence between EDA and other methodologies, such as data mining and resampling. Therefore, conventional conceptual frameworks of EDA might no longer be capable of coping with this trend. In this article, EDA is introduced in the context of data mining and resampling with an emphasis on three goals: cluster detection, variable selection, and pattern recognition. TwoStep clustering, classification trees, and neural networks, which are powerful techniques to accomplish the preceding goals, respectively, are illustrated with concrete examples.

Key words: exploratory data analysis, data mining, resampling, cross-validation, data visualization, clustering, classification trees, neural networks
.

RESUMEN

Hoy por hoy existen diseminadas varias definiciones erróneas acerca del análisis de datos exploratorio (ADE). Una de tales definiciones afirma que ADE es opuesto a la modelación estadística. De hecho, en ADE no se trata de obviar modelaciones y pre-concepciones, al contrario se trata de hacer análisis usando no únicamente pre-concepciones fuertes, lo que en si hace legitimo el uso de modelación en ADE. Además, la naturaleza de ADE ha estado cambiando debido a la emergencia de nuevos métodos y la convergencia de ADE con otras metodologías, tales como la extracción de datos y el remuestreo. Por tanto, las definiciones convencionales de ADE no dan cuenta de su estado actual. En este artículo, ADE se presenta en el contexto de la extracción de datos y el remuestreo haciendo énfasis en tres objetivos: detección de conglomerados, selección de variables, y reconocimiento de patrones. Las técnicas de clasificación en dos pasos, árboles de clasificación, y redes neuronales sirven como ejemplos para lograr los objetivos delineados.
.

Palabras clave: Análisis de datos exploratorio, extracción de datos, remuestreo, validación cruzada, visualización de datos, clasificación, arboles de clasificación, redes neuronales.
.

Artículo recibido/Article received: December 2009, Artículo aceptado/Article accepted: March 15/2010

Dirección correspondencia/Mail Address:
Chong Ho Yu, Ph.D. Director of Research and Assessment. Applied Learning Technologies Institute. Arizona State University 1475 N Scottsdale Rd Scottsdale, AZ 85257.
Email: chonghoyu@yahoo.com; chonghoyu@gmail.com

Exploratory data analysis (EDA) was introduced by Tukey and his colleagues about four decades ago (Tukey, 1969, 1977, 1986a, 1986b, 1986c, Tukey & Wilk, 1986), and since then numerous publications regarding EDA have become available to researchers (e.g. Behrens, 1997; Behrens & Yu, 2003; Fielding, 2007; Martinez, 2005; Myatt, 2007; Schwaiger, & Opitz, 2001; Velleman & Hoaglin, 1981). Although EDA is no longer considered a new methodology, the author of this article, based upon teaching and consulting experiences, observed that today there are still quite a few widespread misconceptions of EDA. This phenomenon is partly due to the fact that EDA is a philosophy or mentality (skepticism and openness) (Hartwig & Dearing, 1979) rather than being a fixed set of formal procedures, and it is also partly owing to the trend that emerging methods, such as data mining and resampling, have been gradually changing the nature of EDA. As a remedy to those misconceptions, this paper will start with clarifying what EDA is not, and then introducing conventional EDA and its limitations. Next, EDA in the new context of data mining and resampling will be illustrated with concrete examples. Although these examples are from education or educational psychology, the principles of analyzing these data sets could be extended to experimental psychology as well as other branches of psychology.

WHAT IS NOT EDA?

When some people claim that their methodology is exploratory, what they actually mean is that they are not sure what they are looking for. Unfortunately, poor research is often implemented in the name of EDA. During data collection, some researchers flood their subjects with hundred of survey items since their research questions are not clearly defined and their variables are not identified. While it is true that EDA does not require a pre-determined hypothesis to be tested, it does not justify the absence of research questions or ill-defined variables.

Another common misperception is that EDA is said to be opposed to statistical modeling. Because EDA is different from confirmatory data analysis (CDA), a set of statistical procedures aiming to confirm a pre-formulated hypothesis using either p-values or confidence intervals, some researchers believe that anything associated with modeling or pre-conceived ideas about the data would disqualify the analysis as a form of EDA. Gelman (2004) found that either EDA is often implemented in the absence of modeling or that EDA is used only in the early stages of model formulation, but disappears from the radar screen after the model is generated. Actually, EDA employs data visualization as a primary tool, which is often used in model diagnostics. For example, a quantile-quantile plot can be drawn to examine the gap between the data and the empirical distribution of a model. Sometimes, data should be explored in an iterative fashion by fitting as much structure as possible into a model and then using graphs to find patterns that represent deviations from the current model (Gelman, 2004). Following this line of reasoning, model-based clustering, which is based upon certain probabilistic inferences, is considered legitimate in EDA (Martinez, 2005).

It is difficult for a data analyst to start with a "blank mind" and explore the data without any reference. Traditionally, researchers classify the modes of reasoning in research as induction (data-driven) and deduction (theory or hypothesis driven). Actually, there is a third avenue: abduction. Abductive reasoning does not necessarily start with fully developed models or no models at all. For example, when Kepler developed his astronomical model, he had some basic preconceptions, which were very general "hunches" about the nature of motion and forces, and also the basic idea that the Sun is the source of the forces driving the planetary system. It is beyond the scope of this article to thoroughly discuss abductive logic. Interested readers are advised to consult Yu (1994, 2006, 2009a). In alignment to abduction, the essence of EDA is not about putting aside all modeling and preconceptions; rather, researchers are urged not to start the analysis with a *strong* preconception only.

CONVENTIONAL VIEWS OF EDA

Exploratory data analysis was named by Tukey (1977) as an alternative to CDA. As mentioned before, EDA is an attitude or philosophy about how data analysis should be carried out, instead of being a fixed set of techniques. Tukey (1977) often related EDA to detective work. In EDA, the role of the researcher is to explore the data in as many ways as possible until a plausible "story" of the data emerges. Therefore, the "data detective" should be skeptical of the "face" value of the data and keep an open mind to unanticipated results when the hidden patterns are unearthed.

Throughout many years, different researchers formulated different definitions, classifications, and taxonomies of EDA. For example, Velleman and Hoaglin (1981) outlined four basic elements of exploratory data analysis: residual, re-expression (data transformation), resistant, and display (data visualization). Based upon Velleman and Hoaglin's framework, Behrens and Yu (2003) elaborated the above four elements with updated techniques, and renamed "display" to "revelation." Each of them is briefly introduced as follows:

1. Residual analysis: EDA follows the formula that data = fit + residual or data = model + error. The fit or the model is the expected values of the data whereas the residual or the error is the values that deviate from that expected value. By examining the residuals, the researcher can assess the model's adequacy (Yu, 2009b).

2. Re-expression or data transformation: When the distribution is skewed or the data structure obscures the

pattern, the data could be rescaled in order to improve interpretability. Typical examples of data transformation include using natural log transformation or inverse probability transformation to normalize a distribution, using square root transformation to stabilize variances, and using logarithmic transformation to linearize a trend (Yu, 2009b).

3. Resistance procedures: Parametric tests are based on the mean estimation, which is sensitive to outliers or skewed distributions. In EDA, resistant estimators are usually used. The following are common examples: median, trimean (a measure of central tendency based on the arithmetic average of the values of the first quartile, the third quartile, and the median counted twice), Winsorized mean (a robust version of the mean in which extreme scores are pulled back to the majority of the data), and trimmed mean (a mean without outliers). It is important to point out that there is a subtle difference between "resistance" and "robustness" though two terms are usually used interchangeably. Resistance is about being immune to outliers while robustness is about being immune to assumption violations. In the former, the goal is to obtain a data summary, while in the latter the goal is to make a probabilistic inference.

4. Revelation or data visualization: Graphing is a powerful tool for revealing hidden patterns and relationships among variables. Typical examples of graphical tools for EDA are Trellis displays and 3D plots (Yu & Stockford, 2003). Although the use of scientific and statistical visualization is fundamental to EDA, they should not be equated, because data visualization is concerned with just one data characterization aspect (patterns) whereas EDA encompasses a wider focus, as introduced in the previous three elements (NIST Semantech, 2006).

According to NIST Semantech (2006), EDA entails a variety of techniques for accomplishing the following tasks: 1) maximize insight; 2) uncover underlying structure; 3) extract important variables; 4) detect outliers and anomalies; 5) test underlying assumptions; 6) develop parsimonious models; and 7) determine optimal factor settings. Comparing the NIST's EDA approach with Velleman and Hoaglin's, and Behrens and Yu's, it is not difficult to see many common threads. For example, "maximize insight" and "uncover underlying structure" is similar to revelation.

LIMITATIONS OF CONVENTIONAL VIEWS TO EDA

Although the preceding EDA framework provides researchers with helpful guidelines in data analysis, some of the above elements are no longer as important as before due to the emergence of new methods and convergence between EDA and other methodologies, such as data mining and resampling. Data mining is a cluster of techniques that has been employed in the Business Intelligence (BI) field for many years (Han & Kamber, 2006). According to Larose (2005), data mining is the process of automatically extracting useful information and relationships from immense quantities of data. Data mining does not start with a strong preconception, a specific question, or a narrow hypothesis, rather it aims to detect patterns that are already present in the data. Similarly, Luan (2002) views data mining as an extension of EDA. Like EDA, resampling departs from theoretical distributions used by CDA. Rather, its inference is based upon repeated sampling within the same sample, and that is why this school is called resampling (Yu, 2003, 2007). How these two methodologies alter the features of EDA will be discussed next.

Checking assumptions

In multiple regression analysis the assumption of the absence of multicollinearity (high correlations among predictors) must be met for the independent variables. If mutlicollinearity exists, probably the variance, standard error, and parameter estimates are all inflated. In addition to computing the variance inflation factor, it is a common practice to use a scatterplot matrix, a data visualization technique for EDA, to examine the inter-relationships among the predictors. While checking underlying assumptions plays an important role in conventional EDA, many new EDA techniques based upon data mining are non-parametric in nature. For example, recursive partition trees and neural networks are immune to multicollinearity (Carpio, & Hermosilla, 2002; Fielding, 2007).

Spotting outliers

In the past it was correct to say that outliers were detrimental to data analysis because the slope of a regression line could be driven by just a single extreme datum point. Thus, it is logical to assert that spotting outliers is an indispensable step in EDA. However, TwoStep clustering, a sophisticated EDA algorithm, has built-in mechanisms to handle outliers during the clustering process. Actually, before the analysis the researcher could not tell which case is an outlier because the references (clusters) have not been made yet. Further, the recursive partition tree, which is a newer EDA technique arising from data mining, is also immune against outliers (Fielding, 2007).

Data transformation

Data transformation is used as a powerful technique to improve the interpretability of the data. But in the recursive partition tree, the independent variables do not require any transformation at all (Fielding, 2007). In addition, Osborne (2002) asserted that many transformations that reduce non-normality by changing the spacing between data points raises issues in the interpretation of data, rather than improving it. If transformations are done correctly, all data points should remain in the same relative order as prior to transformation and this does not affect researchers' interpretations of scores. This might be problematic if the original variables were meant to be interpreted in a straight-forward fashion, such as annual income and age. After the transformations, the new variables might become much more complex to interpret. Even if transformation is needed, some data mining procedures, such as neural networks, perform this task in a hidden layer without the intervention of the analyst.

Transparency and interpretability

Data visualization aims to improve transparency of the analytical process. While hypothesis testers submit the data to complicated algorithms without understanding how the Wilk's Lambda and the p-value are computed, data visualizers could directly "see" the pattern on the graph. Not only do data analysts like the transparency and interpretability that results from visualization, but most teachers and speakers also like to employ graphing techniques to present abstract results and complicated data structures in a concrete and appealing manner (Yu & Stockford, 2003). Interestingly enough, although variable selection is considered an objective of EDA by NIST Sematech (2006) and many other exploratory data analysts, traditional variable selection procedures, such as stepwise regression, are usually excluded from the arena of EDA for lacking visualization and transparency. However, it is important to note that the neural network, another new EDA technique based on data mining, is considered a "black box" because of a lack of transparency in the process (Fielding, 2007). Nevertheless, it is still a powerful tool for pattern recognition.

Resampling and validation

Confirmatory data analysis employs probabilistic inferences and thus the results yielded from CDA are said to posses a high degree of generalizability. In contrast, EDA focuses on pattern recognition using the data at hand. For this reason, EDA is said to aim at hypothesis generation as a complementary approach to CDA (Behrens & Yu, 2003). Traditional EDA techniques might pass the initial findings (suggested factors or hypotheses) to CDA for further inquiry. However, with the use of resampling, new EDA can go beyond the initial sample to validate the finding. This feature will be further discussed in a later section. Moreover, in the past, comparing EDA and CDA results was just like comparing an apple and an orange. For example, EDA does not return a p-value at all. Nevertheless, today some new data mining-based EDA techniques allow the researcher to compare EDA results against those produced from conventional procedures (e.g. regression). How different solutions concur with each other could be viewed as a type of validation.

A NEW EDA FRAMEWORK

Goal-oriented, not means-oriented

Nevertheless, certain conventional EDA elements are still indispensable. For example, in data mining many iterative processes still rely on residual analysis, and no doubt data visualization is essential to examining hidden patterns. But taking all of the above into account, it is obvious that some of the conventional elements of EDA are not fully applicable to the new development. It doesn't necessarily imply that checking assumptions, spotting outliers, transforming data, and so on are obsolete; rather, they could still be useful in some situations. However, there are other EDA procedures for us to use to get around them. Hence, it is time to reconsider the appropriateness of the existing EDA framework. One of the problems of those conventional approaches is that the characteristics of EDA are tied to both the attributes of the data (distribution, variability, linearity, outliers, measurement scales …etc) and the final goals (detecting clusters, screening variables, and unearthing hidden patterns and complex relationships). In fact, dealing with the attributes of the data is just the means instead of the ends, and as demonstrated above, some data characteristics are no longer considered problematic to new EDA. However, if EDA is characterized by a goal-oriented approach, then detecting clusters, screening variables, and unearthing hidden relationships would still be applicable to all techniques no matter what advanced procedures are introduced in the future.

In the following section each of the three goals of EDA stated above will be discussed. There are numerous new EDA techniques belonging to the preceding three categories. Due to space limitations, only one technique will be illustrated in each category. In addition, because variable selection and pattern recognition methods are guided by a response variable, they are considered "supervised learning methods." On the other hand, clustering techniques have no dependent variable as a reference, and thus they are called "unsupervised learning methods." "Learning" in this context means these approaches are data-driven i.e. the algorithms learn from the data.

CATEGORIES AND TECHNIQUES OF EDA

Clustering: TwoStep cluster analysis

Clustering is essentially grouping observations based upon their proximity to each other on multiple dimensions. At first glance, clustering analysis is similar to discriminant analysis. But in the latter the analyst must know the group membership for the classification in advance. Because discriminant analysis assigns cases to pre-existing groups, it is not as exploratory as cluster analysis, which aims to identify the grouping categories in the first place.

If there are just two dimensions (variables), the analyst could simply use a scatterplot to look for the clumps. But when there are many variables, the task becomes more challenging and thus it necessitates algorithms. There are three major types of clustering algorithms: 1) Hierarchical clustering, 2) non-hierarchical clustering (k-mean clustering), and 3) TwoStep clustering. The last one is considered the most versatile because it has several desirable features that are absent in other clustering methods. For example, both hierarchical clustering and k-mean clustering could handle continuous variables only, but TwoStep clustering accepts both categorical and continuous variables. This is the case because in TwoStep clustering the distance measurement is based on the log-likelihood method (Chiu et al., 2001). In computing log-likelihood, the continuous variables are assumed to have a normal distribution and the categorical variables are assumed to have a multinomial distribution. Nevertheless, the algorithm is reasonably robust against the violation of these assumptions, and thus assumption checking is unnecessary. Second, while k-mean clustering requires a pre-specified number of clusters and therefore strong prior knowledge is required, TwoStep clustering is truly data-driven due to its capability of automatically returning the number of clusters. Last but not least, while hierarchical clustering is suitable to a small data set only, TwoStep clustering is so scalable that it could analyze thousands of observations efficiently.

As the name implies, TwoStep clustering is composed of two steps. The first step is called preclustering. In this step, the procedure constructs a cluster features (CF) tree by scanning all cases one by one (Zhang et al., 1996). When a case is scanned, the pre-cluster algorithm applies the log likelihood distance measure to determine whether the case should be merged with other cases or form a new precluster on its own and wait for similar cases in further scanning. After all cases are exhausted, all preclusters are treated as entities and become the raw data for the next step. In this way, the task is manageable no matter how large the sample size is, because the size of the distance matrix is dependent on just a few preclusters rather than all cases. Also, the researcher has the option to turn on outlier handling. If this option is selected, entries that cannot fit into any preclusters are treated as outliers at the end of CF-tree building. Further, in this preclustering step, all continuous variables are automatically standardized. In other words, there is no need for the analyst to perform outliers detection and data transformation in separate steps.

In step two, the hierarchical clustering algorithm is applied to the preclusters and then propose a set of solutions. To determine the best number of clusters, each solution is compared against each other based upon the Akaike Information Criterion (AIC) (Akaike, 1973) or the Bayesian Information Criterion (BIC) (Schwarz, 1978). AIC is a fitness index for trading off the complexity of a model against how well the model fits the data. To reach a balance between fitness and parsimony, AIC not only rewards goodness of fit, but also gives a penalty to over-fitting and complexity. Hence, the best model is the one with the lowest AIC value. However, both Berk (2008) and Shmueli (2009) agreed that although AIC is a good measure of predictive accuracy, it can be over-optimistic in estimating fitness. In addition, because AIC aims to yield a predictive model, using AIC for model selection is inappropriate for a model of causal explanation. BIC was developed as a remedy to AIC. Like AIC, BIC also uses a penalty against complexity, but this penalty is much stronger than that of the AIC. In this sense, BIC is in alignment to Ockham's razor: Given all things being equal, the simplest model tends to be the best one.

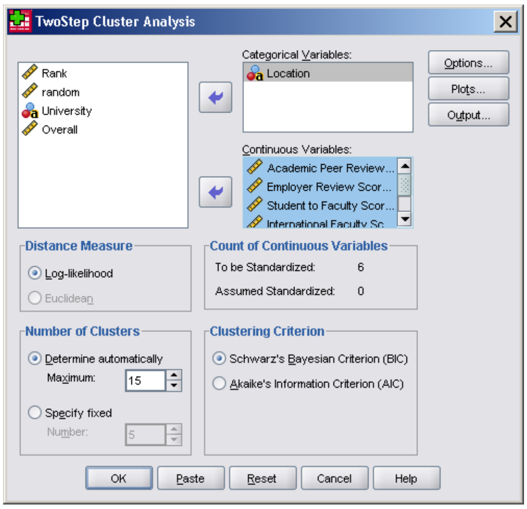
To illustrate TwoStep clustering, a data set listing 400 of the world's best colleges and universities compiled by *US News and World Report* (2009) was utilized. The criteria used by *US News and World Report* for selecting the best institutions include: Academic peer review score, employer review score, student to faculty score, international faculty score, international students score, and citations per faculty score. However, an educational researcher might not find the list helpful because the report ranks these institutions by the overall scores. It is tempting for the educational researcher to learn about how these best institutions relate to each other and what their common threads are. In addition to the preceding measures, geographical location could be taken into account.

Because the data set contains both categorical and continuous variables, the researcher employed the TwoStep clustering analysis in Predictive Analytical Software (PASW) Statistics (SPSS Inc., 2009). It is important to note that the clustering result may be affected by the order of the cases in the file. In the original data set, the table has been sorted by the rank in an ascending order. In an effort to minimize the order effect, the cases were re-arranged in random order before the analysis was conducted. To run a TwoStep cluster analysis, the researcher must assign the categorical and continuous variables into the proper fields, as shown in Figure 1, using BIC instead of AIC for simplicity.

In this analysis a three-cluster solution is suggested (see Figure 2). Cluster 1 is composed of all European institutions, whereas Cluster 2 includes colleges and universities in Australia, New Zealand, Asia, and Africa. Cluster 3 consists of North American and South

American institutions. Other characteristics of these clusters will be discussed next.

Figure 1. *Options in TwoStep cluster analysis*



TwoStep Cluster Analysis

Rank
random
University
Overall

Categorical Variables:
Location

Continuous Variables:
Academic Peer Review...
Employer Review Scor...
Student to Faculty Scor...
International Faculty Sc...

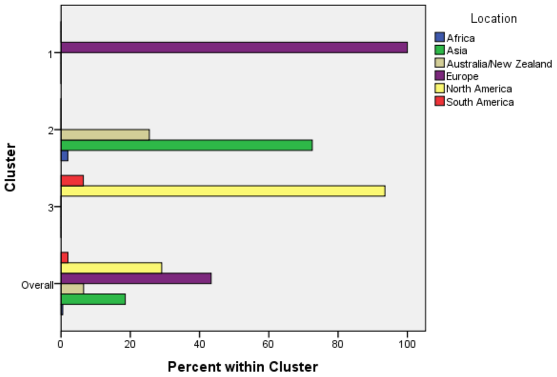
Distance Measure
● Log-likelihood
○ Euclidean

Count of Continuous Variables
To be Standardized: 6
Assumed Standardized: 0

Number of Clusters
● Determine automatically
Maximum: 15
○ Specify fixed
Number: 5

Clustering Criterion
● Schwarz's Bayesian Criterion (BIC)
○ Akaike's Information Criterion (AIC)

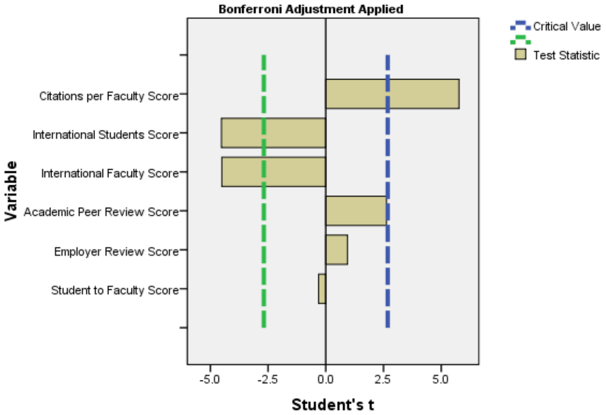
Figure 2. *Barchart of within cluster percentage*



PASW returns many tables and graphs for the analyst to examine the results. Due to space constraints, only a few will be discussed here. For example, in Cluster 3 three variables are considered important to distinguishing Cluster 3 from the other two clusters. The three important variables are citations per faculty score, international students score, and international faculty score, because their t-statistics exceed the critical value (see Figure 3).

The attributes of each cluster could be further examined using the centroids table (Table 1). Cluster 1 is characterized by high international students score, high international faculty score, and moderate citations per faculty score. On the other hand, Cluster 2 possesses the following characteristics: moderate international students score, moderate international faculty score, and low citations per faculty score. In the last cluster, both international faculty score and international student score are the lowest, but its citations per faculty score is the best.

Figure 3. *Importance of variables for setting clusters apart*

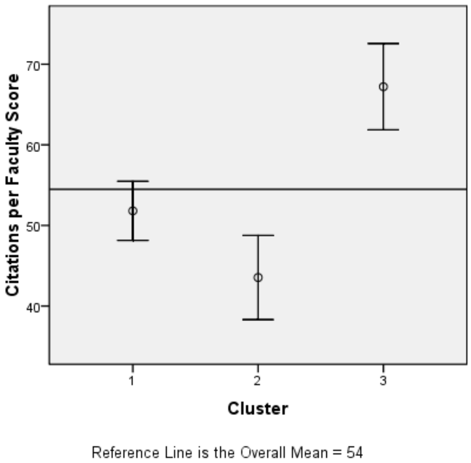


The 95% confidence intervals of citations per faculty score clearly indicate that Cluster 3 substantively outperforms the two other clusters (see Figure 4). Actually, in Cluster 3 most institutions are located in the US, and this implies that although the best American universities are successful in research in terms of citations and recognition, they lack a strong international component comparing with their overseas counterparts. At the end, the researcher labels the three clusters as follows: 1) Cluster 1: International-emphasis institutions; 2) Cluster 3: Research-emphasis institutions; and 3) Cluster 2: Balanced (between international-emphasis and research-emphasis) institutions.

Table 1. *Centroids table.*

Centroids					
		Cluster			
		1	2	3	Combined
Academic Peer Review Score	Mean	54.13	58.86	64.40	58.53
	Std. Deviation	21.512	24.352	24.860	23.678
Employer Review Score	Mean	52.97	62.40	59.72	57.48
	Std. Deviation	26.972	23.897	26.545	26.338
Student to Faculty Score	Mean	55.24	51.90	52.94	53.67
	Std. Deviation	25.418	24.268	26.305	25.388
International Faculty Score	Mean	**59.90**	**50.10**	**43.68**	52.35
	Std. Deviation	26.841	31.734	21.432	27.538
International Students Score	Mean	**62.66**	**46.22**	**43.85**	52.61
	Std. Deviation	25.117	31.604	21.571	27.354
Citations per Faculty Score	Mean	**51.81**	**43.54**	**67.21**	54.48
	Std. Deviation	19.975	21.671	24.554	23.711

Figure 4. *95% confidence intervals*



Variable selection: Recursive partition trees

Classification trees, developed by Breiman et al. (1984), aim to find which independent variable(s) can successfully make a decisive split of the data by dividing the original group of data into pairs of subgroups in the dependent variable. Because classification trees can provide guidelines for decision-making, they are also known as decision trees. In addition, because at each decision point the data are partitioned and each partition is further partitioned independently of all other partitioned data until all possible splits are exhausted, they are also called recursive partition trees (Fielding, 2007).

In programming terminology, a classification tree can be viewed as a set of "nested-if" logical statements. Breiman et al. used the following example for illustrating nested-if logic. When heart attack patients are admitted to a hospital, three pieces of information are most relevant to the survival of patients: What is the patient's minimum systolic blood pressure over the initial 24 hour period? What is his/her age? Does he/she display sinus tachycardia? The answers to these three questions can help the doctor to make a quick decision: "If the patient's minimum systolic blood pressure over the initial 24 hour period is greater than 91, then if the patient's age is over 62.5 years, then if the patient displays sinus tachycardia, then and only then the patient is predicted not to survive for at least 30 days." These nested-if decisions can be translated into a graphical form as a tree structure.

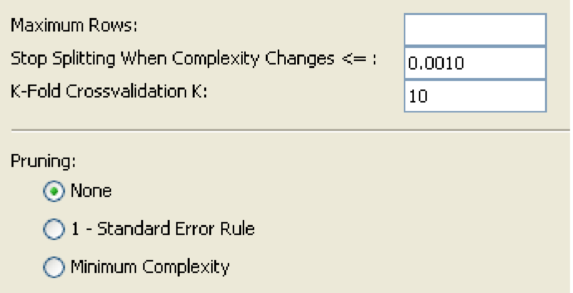
As mentioned before, classification trees can accept the original data without transformation, regardless of the distribution and scaling. Specifically, the algorithm is invariant to monotonic transformation that retains the rank order of the observations. Thus, making a logarithmic transformation of data will lead to the same result. Additionally, classification trees are robust against outliers, because the data set is partitioned into many nodes during the exploratory process, and as a result, the effect of outliers is confined into their own nodes. In other words, those outliers have no effects on other nodes and the efficacy of the overall result (Fielding, 2007).

Like many other data mining procedures, classification trees employed cross-validation, (Krus & Fuller, 1982), which is a form of resampling, to enhance its predictive power. Put simply, cross-validation divides the data set into training sets and testing sets. Exploratory modeling using the training data set inevitably tends to overfit the data. But in the subsequent modeling using the testing data set, the overfitted model will be revised in order to enhance its generalizability. It is better to overfit the model and then scale back to the optimal point. If a model is built from a forward stepping approach and ended by a stopping rule, the researcher will miss the opportunities of seeing what might be possible and better ahead (Quinlan, 1993).

In the following discussion, the data set "Programme for International Student Assessment" (PISA) was utilized to illustrate classification trees. PISA is a series of assessments in science, mathematics, and reading. It is sponsored by the Organization for Economic Cooperation and Development (OECD, 2006), and administered internationally to 15-year-olds from different countries. In addition to test scores, PISA also administers many other instruments, such as the cognitive item test, the school questionnaire, the student demographic questionnaire, and the information and communication technology familiarity component for students questionnaire. In this example, using the US and Canadian observations (n=22,601), the researcher would like to find out which variables could best predict performance in the science test. While the researcher was burdened with hundreds of variables listed in all preceding instruments, he turned to classification trees in Spotfire Miner (TIBCO, 2009).

Using the logit yielded form an Item Response Theory (IRT) analysis, students were divided into high and low performers, with this grouping variable became the outcome variable. To run a classification tree, the researcher simply entered the dependent variable (performance in terms of logit) and all potential predictors. As mentioned before, outlier detection, data transformation, and assumption check are not needed. Classification trees have built-in cross-validation (resampling) mechanisms. But it is important to note that by default Spotfire sets the K-fold cross-validation K to "0." It is advisable to change it to 2 or more. Kohavi (1995) suggested that 10-fold partitioning produced the best result, however, "10" is not the magic number. Thus, the researcher could try out different settings. Also, It is tempting to use some stopping rule to prune the tree and "minimum complexity" might be attractive to researchers that favor a simple model (see Figure 5). But it is better to select "none" for pruning because as mentioned before, premature stopping disallows the researcher to see what is possible and better.

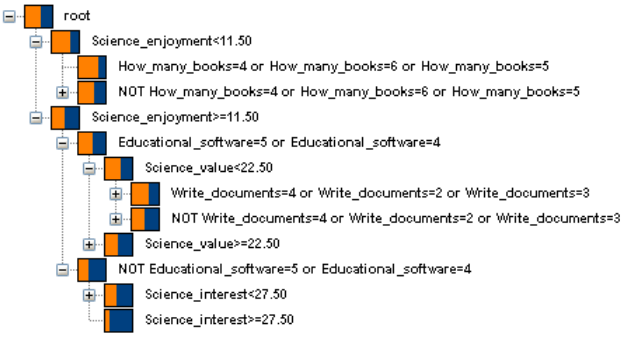
Figure 5. *K-fold cross-validation and pruning criterion.*



Maximum Rows:
Stop Splitting When Complexity Changes <= : 0.0010
K-Fold Crossvalidation K: 10

Pruning:
● None
○ 1 - Standard Error Rule
○ Minimum Complexity

Figure 6. *Optimal partition tree after pruning.*



```
⊟ root
  ⊟ Science_enjoyment<11.50
      How_many_books=4 or How_many_books=6 or How_many_books=5
    ⊞ NOT How_many_books=4 or How_many_books=6 or How_many_books=5
  ⊟ Science_enjoyment>=11.50
    ⊟ Educational_software=5 or Educational_software=4
      ⊟ Science_value<22.50
        ⊞ Write_documents=4 or Write_documents=2 or Write_documents=3
        ⊞ NOT Write_documents=4 or Write_documents=2 or Write_documents=3
      ⊞ Science_value>=22.50
    ⊟ NOT Educational_software=5 or Educational_software=4
      ⊞ Science_interest<27.50
        Science_interest>=27.50
```

After the job was submitted, Spotfire returned a suggested tree model, as shown in Figure 6, the orange portion of each rectangle depicts high performers while the blue portion signifies weaker performers. The classification tree identified science enjoyment, the number of books at home, frequent use of educational software, frequent use of computers for writing documents, science interest, and science value as the most important predictor to performance in the PISA science test. This model is considered optimal because when the tree grows by further partitioning, these variables keep recurring. In other words, increasing complexity does not yield additionally useful information, and thus the redundant components were manually pruned.

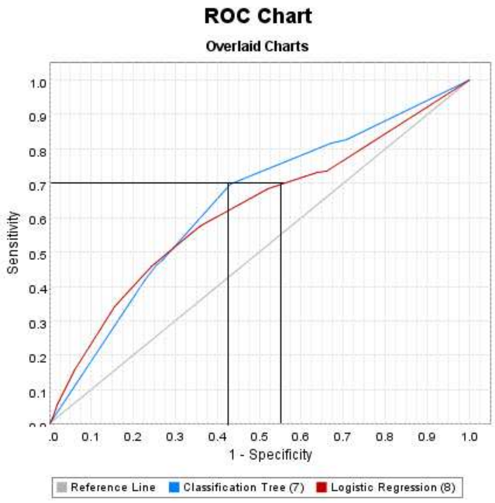
A logistic regression model was run side by side with the preceding classification tree. Unlike its classification tree counterpart, the logistic regression model suggested a longer list of important predictors: Science enjoyment, the number of books at home, frequent use of educational software, frequent use of computers for writing documents, frequent use of computers for writing programs, frequent use of computers for downloading music, the number of TV sets, frequent use of spreadsheets, frequent use of computers for playing games, the number of computers at home, frequent use of graphics programs, frequent use of computers for online communication, frequent use of computers for collaborating on the Internet. However, when there are too many predictors, the reliability of the parameter estimates decrease (Fielding, 2007). The predictive power of the two approaches was evaluated by both classification agreement and ROC curves. Table 2 indicates that the classification tree outperforms the logistic regression model in predicting both high (1) and weaker performers (0).

Table 2. *Classification agreement between the predicted and observed for all students.*

	Predicted and observe matched (1)	Predicted and observe matched (0)	Overall
Classification tree	84.1%	40.4%	67.00%
Logistic regression	83.9%	39.0%	65.9%

This assessment is bolstered by the overlaid ROC curves, which illustrate sensitivity (true positive rate) and 1 – specificity (false positive rate). The ideal prediction outcomes are 100% sensitivity (all true positives are found) and 100% specificity (no false positives are found). In Figure 7, the 45 degree diagonal gray line represents the baseline. When there is no modeling, the probability is .5. Thus, a good classifier should depict a ROC curve leaning towards the upper left of the graph. Figure 6 shows that overall the classification tree, shown by a blue line, is superior to the logistic regression, presented by a red line. Specifically, while attempting to achieve the highest true positive rate, the logistic regression modeling is more liberal than its decision tree counterpart. In other words, it makes positive classification with weak evidence and tends to get positive cases correct at the expense of a high false positive rate. For example, when the true positive rate of the logistic regression is .7, its false positive rate is as high as .55. But when decision tree reaches the same true positive rate, its false positive rate is just .425. It is true that in the lower left of the chart (lower true positive rate < .5) the logistic regression is more conservative than the classification tree, but in that area the difference between the two models in terms of the false positive rate is narrow. In summary, no matter whether simplicity, classification agreement or ROC curves was used as the criterion for determining the model choice, it is obvious that the classification tree approach is more advantageous than logistic regression.

Figure 7. *ROC comparing classification tree and logistic regression.*

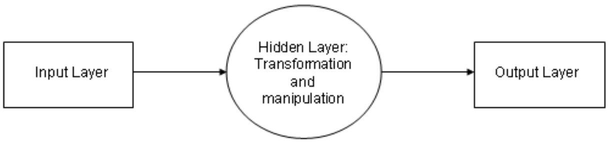


Pattern recognition: Neural networks

While classification trees aim to identify predictors, neural networks can be used for both selecting variable and examining complex relationships (Gonzalez & DesJardins, 2002). Neural networks, as the name implies, try to mimic interconnected neurons in animal brains in order to make the algorithm capable of complex learning for extracting patterns and detecting trends (Kuan, 1994; McMenamin, 1997). Because this approach artificially mimics human neurons in computers, it is also named artificial neural networks. It is built upon the premise that real world data structures are complex and nonlinear, and thus it necessitates complex learning systems. Unlike regression modeling that assumes linearity, neural networks could model linearity and thus they typically outperformed regression (Somers & Casal, 2009).

A trained neural network can be viewed as an "expert" in the category of information it has been given to analyze. This expert system can provide projections given new solutions to a problem and answer "what if" questions. A typical neural network is composed of three types of layers, namely, the input layer, hidden layer, and output layer (see Figure 8). It is important to note that there are three types of layers, not three layers, in the network. There may be more than one hidden layer and it depends on how complex the researcher wants the model to be. The input layer contains the input data; the output layer is the result whereas the hidden layer performs data transformation and manipulation.

Figure 8. *Three layers of a typical neural network*

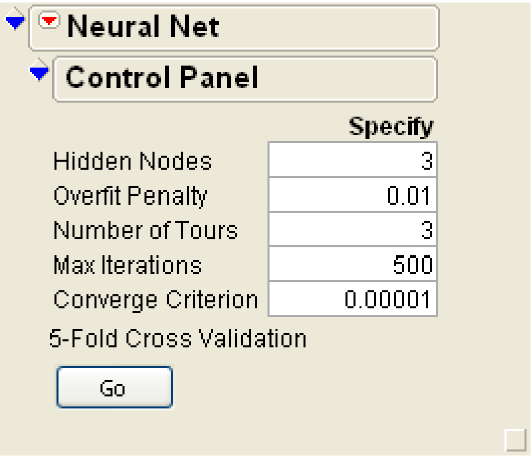


As mentioned in the third section, preliminary data transformation is unnecessary in many data mining techniques, including neural networks. In traditional linear regression the researcher might try different transformation of the predictors, interactions between predictors, or both (e.g. using centered scores for interaction terms). But in neural networks these are automatically processed in the hidden layer. In this sense, linear regression and logistic regression can be viewed as special cases of neural networks that omit the hidden layer (Shmueli, Patel, & Bruce, 2007; Yu, 2009c). Because the input and the output are mediated by the hidden layer that is not transparent to the analyst, neural networks are commonly seen as a "black box."

The network is completely connected in the sense that each node in the layer is connected to each node in the next layer. Each connection has a weight at the initial stage and these weights are just randomly assigned. A common technique in neural networks to fit a model is called back propagation. During the process of back propagation, the residuals between the predicated and the actual errors in the initial model are fed back to the network. In this sense, back propagation is in a similar vein to residual analysis in conventional EDA (Behrens & Yu, 2003). Since the network performs problem-solving through learning by examples, its operation can be unpredictable. Thus, this iterative loop continues one layer at a time until the errors are minimized. Neural networks use multiple paths for model construction. Each path-searching process is called a "tour" and the desired result is that only one best model emerges out of many tours. Like other data mining techniques, neural networks also incorporate cross-validation to avoid capitalization on chance alone in one single sample.

In the following illustration, the example data set was compiled by tracking the continuous enrollment or withdrawal of 6690 sophomore students enrolled at a US university starting in 2003. The dependent variable is a dichotomous variable, retention. In this study, retention is defined as persisting enrollment within the given time frame (2003-2004 academic years, excluding summer). There are three sets of potential predictors: 1) Demographic: This set of predictors includes gender, ethnic, residence (in state/out of state), and location (living on campus/off campus). 2) Pre-college or external academic performance indicators: This set of variables includes high school GPA, high school class rank, SAT scores, ACT scores, transferred hours, and university mathematics placement test scores. 3) Online class hours as a percentage of total hours during the sophomore year. Like the PISA data set, this data set contains so many variables that using CDA might be difficult. Hence, the researcher turned to neural networks for exploring the inter-relationships among these variables.

Figure 9. *Dialog box of neural networks in JMP*



	Specify
Hidden Nodes	3
Overfit Penalty	0.01
Number of Tours	3
Max Iterations	500
Converge Criterion	0.00001

5-Fold Cross Validation

For this analysis, neural networks in JMP (SAS Institute, 2009) were utilized. The very essence of EDA is the freedom of exploration; there is no single best approach. Thus, the researcher could freely enter the numbers of hidden nodes, tours, maximum iterations, and folds of cross-validation, as shown in the following dialog box (Figure 9). After several trials with different settings, the researcher could select the most interpretable one out of a set of suggested results.

Taking clarity of interpretation as the major criterion, the results of the neural net using three hidden layers, three tours, and 5-fold cross-validation are retained for the following discussion. A neural network allows the analyst to examine all possible interactions (see Figure 10). On the right panel of the graph, each rectangle contains the value range of the variable from the lowest to the highest. Inside each rectangle there is a slider for manipulation. When the value of the variable changes, there is a corresponding change in the graph. The analyst can use the slider to superimpose a value grid on the graph and at the same time the rightmost cell shows the exact value of the variable of interest. It is crucial to emphasize that *these are not regular 3-D plots that are commonly found in most EDA packages*, in which frequencies or raw values are usually plotted. Rather, the probabilities on the Z-axis result from adaptive learning through iterative loops.

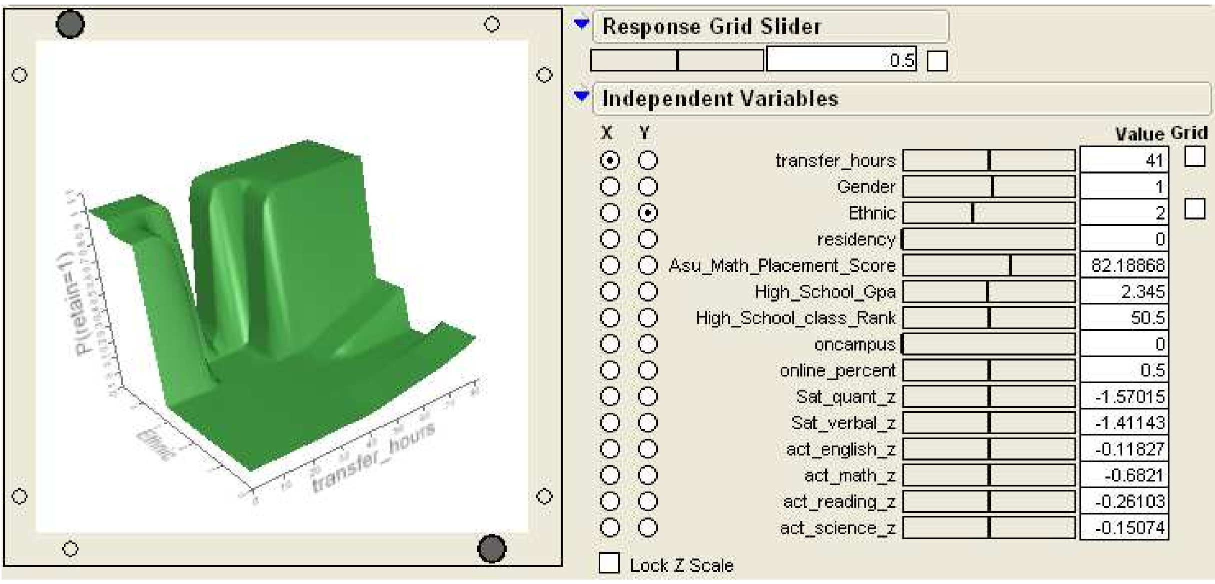
The neural net indicates that the interaction effect between these students is complicated and non-linear. The Y-axis (vertical) of Figure 10 represents the predicted probability of retention, the X-axis denotes the number of transferred hours, and the Z-axis depicts ethnic groups coded as: White = 1, Asian = 2, Hispanic = 3, Black = 4, and Native American = 5. For White and Hispanic students, as the number of transferred hours increases, the probability of retention slightly increases, which is indicated by the gradual slope on the outmost right. For Asian students, an increase in the number of transferred hours does not affect retention rate at all. However, for Black and Native American students, when the amount of transferred hours is low, the probability of continuing enrollment is still high. But there is a sharp drop in probability of retention for Native Americans when the number of transferred credits is between 19 and 31. For Black students, the sudden depression of probability happens between 18 and 38 transferred hours. Afterwards, the probability rises along with the transferred hours.

The interaction between residency and transferred hours is another noteworthy phenomenon. While the probability of retention for non-residents slightly increases as the number of transferred hours increases, the probability for retention climbs up sharply after 42 transferred hours (see Figure 11). It is important to note that 42 is by no means the "magic" cutoff. This may vary from sample to sample, and even from population to population. The main point is that there exists an interaction effect between transferred hours and residency.

EDA AND RESAMPLING

At first glance, exploratory data mining is very similar to conventional EDA except that the former employs certain advanced algorithms for automation. Actually, the differences between conventional EDA and exploratory data mining could be found at the epistemological level. As mentioned before, EDA suggests variables, constructs and hypotheses that are worth pursuing and CDA takes the next step to confirm the findings. However, using resampling (Yu, 2003, 2007), data mining is capable of suggesting and validating a model at the same time. One may argue that data mining should be classified as a form of CDA when validation has taken place. It is important to point out that usually exploratory data mining aims to yield predication rather than theoretical explanations of the relationships between variables (Shmueli & Koppius, 2008; Yu, in press). Hence, the researcher still has to construct a theoretical model in the context of CDA (e.g. structural equation modeling) if explanation is the research objective.

Figure 10. *Interaction between ethnic groups and transferred hours.*



Response Grid Slider

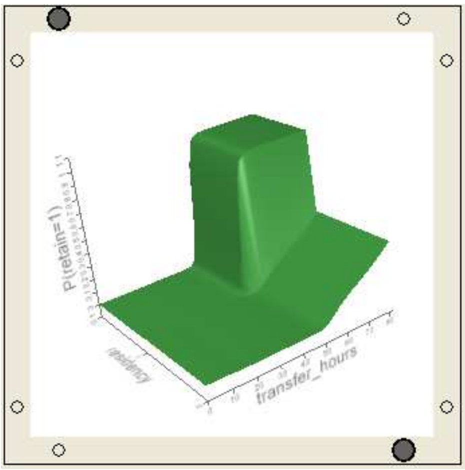
0.5 □

Independent Variables

X	Y		Value	Grid
●	○	transfer_hours	41	□
○	○	Gender	1	
○	●	Ethnic	2	□
○	○	residency	0	
○	○	Asu_Math_Placement_Score	82.18868	
○	○	High_School_Gpa	2.345	
○	○	High_School_class_Rank	50.5	
○	○	oncampus	0	
○	○	online_percent	0.5	
○	○	Sat_quant_z	-1.57015	
○	○	Sat_verbal_z	-1.41143	
○	○	act_english_z	-0.11827	
○	○	act_math_z	-0.6821	
○	○	act_reading_z	-0.26103	
○	○	act_science_z	-0.15074	

□ Lock Z Scale

Figure 11. *Interaction between residency and transferred hours.*



Resampling in the context of exploratory data mining addresses two important issues, namely, generalization across samples and under-determination of theory by evidence (Kieseppa, 2001). It is very common that in one sample a set of best predictors was yielded from regression analysis, but in another sample a different set of best predictors was found (Thompson, 1995). In other words, this kind of model can provide a post hoc model for an existing sample (in-sample forecasting), but cannot be useful in out-of-sample forecasting. This occurs when a specific model is overfitted to a specific data set and thus it weakens generalizability of the conclusion. Further, even if a researcher found the so-called best fit model, there may be numerous possible models to fit the same data.

To counteract the preceding problems, most data mining procedures employed cross-validation to enhance generalizability. For example, to remediate the problem of under-determination of theory by data, neural networks exhaust different models by the genetic algorithm, which begins by randomly generating pools of equations. These initial randomly generated equations are estimated to the training data set and prediction accuracy of the outcome measure is assessed using the test set to identify a family of the fittest models. Next, these equations are hybridized or randomly recombined to create the next generation of equations. Parameters from the surviving population of equations may be combined or excluded to form new equations as if they were genetic traits inherited from their "parents." This process continues until no further improvement in predicting the outcome measure of the test set can be achieved (Baker & Richards, 1999). In addition to cross-validation, bootstrapping, another resampling technique, is also widely employed in data mining (Salford Systems, 2009), but it is beyond the scope of this article to introduce bootstrapping. Interested readers are encouraged to consult Yu (2003, 2007).

CONCLUDING REMARKS

This article introduces several new EDA tools, including TwoStep clustering, recursive classification trees, and neural networks, in the context of data mining and resampling, but these are just a fraction of the plethora of exploratory data mining tools. In each category of EDA there are different methods to accomplish the same goal, and each method has numerous options (e.g. the number of k-fold cross-validation). In evaluating the efficacy of classification trees and other classifers, Wolpert and Macready (1997) found that there is no single best method and they termed this phenomenon “no free lunch” – every output comes with a price (drawback). For instance, simplicity is obtained at the expense of fitness, and vice versa. As illustrated before, sometimes simplicity could be an epistemologically sound criterion for selecting the “best” solution. In the example of PISA data, the classification tree model is preferable to the logistic regression model because of predictive accuracy. And also in the example of world’s best universities, BIC, which tends to introduce heavy penalties to complexity, is more favorable than AIC. But in the example of the retention study, when the researcher suspected that there are entangled relationships among variables, a complex, nonlinear neural net was constructed even though this black box lacks transparency. In one way or the other the data explorer must pay a price. Ultimately, whether a simple and complex approach should be adopted is tied to usefulness. Altman and Royston (2000) asserted that “usefulness is determined by how well a model works in practice, not by how many zeros there are in associated p values” (p.454). While this statement pinpoints the blind faith to p values in using inferential statistics, it is also applicable to EDA. A data explorer should not hop around solutions and refuse to commit himself/herself to a conclusion in the name of exploration; rather, he/she should contemplate about which solution could yield more implications for the research community.

Last but not least, exploratory data mining techniques could be simultaneously or sequentially employed. For example, because both neural networks and classification trees are capable of selecting important predictors, they could be run side by side and evaluated by classification agreement and ROC curves. On other occasions, a sequential approach might be more appropriate. For instance, if the researcher suspects that the observations are too heterogeneous to form a single population, clustering could be conducted to divide the sample into sub-samples. Next, variable selection procedures could be run to narrow down the predictor list for each sub-sample. Last, the researcher could focus on the inter-relationships among just a few variables using pattern recognition methods. The combinations and possibilities are virtually limitless. Data detectives are encouraged to explore the data with skepticism and openness.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *International Symposium on Information Theory* (pp.267–81). Budapest: Akademia Kiado.

Altman, D. G., & Royston, P. (2000).What do we mean by validating a prognostic model? *Statistics in Medicine, 19*, 453-473.

Baker, B. D., & Richards, C. E. (1999). A comparison of conventional linear regression methods and neural networks for forecasting educational spending. *Economics of Education Review, 18*, 405-415.

Behrens, J. T. & Yu, C. H. (2003). Exploratory data analysis. In J. A. Schinka & W. F. Velicer, (Eds.), *Handbook of psychology Volume 2: Research methods in Psychology* (pp. 33-64). New Jersey: John Wiley & Sons, Inc.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*, 131-160.

Berk, R. A. (2008). *Statistical learning from a regression perspective*. New York: Springer.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth International Group.

Carpio, K.J.E. & Hermosilla, A.Y. (2002), On multicollinearity and artificial neural networks, *Complexity International, 10*, Retrieved October 8, 2009, from http://www.complexity.org.au/ci/vol10/hermos01/.

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA 263-268.

Fielding, A. H. (2007). *Cluster and classification techniques for the biosciences*. New York, NY: Cambridge University Press.

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics, 13*, 755-779.

Gonzalez, J., & DesJardins, S. (2002). Artificial neural networks: A new approach to predicting application behavior. *Research in Higher Education, 43*, 235-258.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). Boston, MA: Elsevier.

Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis*. Beverly Hills, CA: Sage Publications.

Kieseppa, I. A. (2001). Statistical model selection criteria and the philosophical problem of underdetermination. *British Journal for the Philosophy of Science, 52*, 761–794.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In C. S. Melish (Ed.), *Proceedings of the 14th International Joint Conference on Artificial*

Intelligence (pp. 1137-1143). San Francisco, CA: Morgan Kauffmann.

Krus, D. J. & Fuller, E. A. (1982). Computer-assisted multicross-validation in regression analysis. *Educational and Psychological Measurement, 42*, 187-193.

Kuan, C., & White, H. (1994). Artificial neural networks: An econometric perspective. *Econometric reviews, 13*, 1-91.

Larose, D. (2005). *Discovering knowledge in data: An introduction to data mining*. NJ: Wiley-Interscience.

Luan, J. (2002). Data mining and its applications in higher education. In A. Serban & J. Luan (Eds.), *Knowledge management: Building a competitive advantage in higher education* (pp. 17-36). PA: Josey-Bass.

Martinez, W. L. (2005). *Exploratory data analysis with MATLAB*. London: Chapman & Hall/CRC.

McMenamin, J. S. (1997). A primer on neural networks for forecasting. *Journal of Business Forecasting, 16*, 17–22.

Myatt, G. (2007). *Making sense of data: A practical guide to exploratory data analysis*. Hoboken, NJ: John Wiley & Sons.

NIST Sematech. (2006). *What is EDA?* Retrieved September 30, 2009, from http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

Organization for Economic Cooperation and Development. (2006). *Programme for international student assessment*. Retrieved July 2, 2009, from http://www.oecd.org/pages/0,3417,en_32252351_32235731_1_1_1_1_1,00.html

Osborne, J. (2002). Notes on the use of data transformations. *Practical Assessment, Research & Evaluation, 8*(6). Retrieved September 30, 2009 from http://PAREonline.net/getvn.asp?v=8&n=6.

Quinlan, J. R. (1993). *C4.5 programs for machine learning.* San Francisco, CA: Morgan Kaufmann.

Salford Systems. (2009). Random forest. [Computer software and manual]. San Diego, CA: Author.

SAS Institute. (2007). JMP 8 [Computer software and manual]. Cary, NC: Author.

Schwaiger, M., & Opitz, O. (Eds.). (2001). *Exploratory data analysis in empirical analysis*. New York: Springer.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Shmueli, G. (2009). *To explain or to predict?* Retrieved March 1, 2009, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1351252

Shmueli, G., & Koppius, O. (2008) *Contrasting predictive and explanatory modeling in IS research*. Robert H. Smith School Research Paper No. RHS 06-058. Retrieved August 21, 2009, from http://ssrn.com/abstract=1112893

Shmueli, G., Patel, N., & Bruce, P. (2007). *Data mining for business intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner*. Hoboken, N.J.: Wiley-Interscience.

Somers, M. J., & Casal, J. C. (2009). Using artificial neural networks to model nonlinearity: The case of the job satisfaction–job performance relationship. *Organizational Research Methods, 12*, 403-417.

SPSS, Inc. (2009). PASW Statistics 17 [Computer software and manual]. Chicago, IL: Author.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55*, 525-534.

TIBCO (2009). Spotfire Miner [Computer software and manual]. Palo Alto, CA: Author.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*, 83-91.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tukey, J. W (1986a). Data analysis, computation and mathematics. In L. V. Jones (Ed.), *The collected works of John W. Tukey: Vol. IV. Philosophy and principles of data analysis: 1965-1986* (pp. 753-775). Pacific Grove, CA: Wadsworth. (Original work published 1972).

Tukey, J. W (1986b). Exploratory Data Analysis as part of a larger whole. In L. V. Jones (Ed.), *The collected works of John W. Tukey: Vol. IV. Philosophy and principles of data analysis: 1965-1986* (pp. 793-803). Pacific Grove, CA: Wadsworth. (Original work published 1973).

Tukey, J. W (1986c). The future of data analysis. In L. V. Jones (Ed.), *The collected works of John W. Tukey: Vol. III. Philosophy and principles of data analysis: 1949-1964* (pp. 391-484). Pacific Grove, CA: Wadsworth. (Original work published 1962).

Tukey, J. W., & Wilk, M. B. (1986). Data analysis and statistics: An expository overview. In L. V. Jones (Ed.), *The collected works of John W. Tukey: Vol. IV. Philosophy and principles of data analysis: 1965-1986* (pp. 549-578). Pacific Grove, CA: Wadsworth. (Original work published 1966).

US News and World Report. (2009, June 18). *World's best colleges and universities*. Retrieved October 5, 2009, from http://www.usnews.com/articles/education/worlds-best-colleges/2009/06/18/worlds-best-colleges-top-400.html

Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Boston, MA: Duxbury Press

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67–82.

Yu, C. H. (1994, April). *Induction? Deduction? Abduction? Is there a logic of EDA?* Paper presented at the Annual Meeting of American Educational Research Association, New Orleans, LA (ERIC Document Reproduction Service No. ED 376 173).

Yu, C. H. (2003). Resampling methods: Concepts, applications, and justification. *Practical Assessment Research and Evaluation, 8*(19). Retrieved July 4, 2009, from http://pareonline.net/getvn.asp?v=8&n=19

Yu, C. H. (2006). *Philosophical foundations of quantitative research methodology*. Lanham, MD: University Press of America.

Yu, C. H. (2007). Resampling: A conceptual and procedural introduction. In Jason Osborne (Ed.), *Best practices in quantitative methods* (pp. 283-298). Thousand Oaks, CA: Sage Publications.

Yu, C. H. (2009a). *Causal inferences and abductive reasoning: Between automated data mining and latent constructs*. Saarbrücken, Germany: VDM-Verlag.

Yu, C. H. (2009b). *Exploratory data analysis and data visualization*. Retrieved October 10, 2009, from http://www.creative-wisdom.com/teaching/WBI/EDA.shtml

Yu, C. H. (2009c). *Multi-collinearity, variance Inflation and orthogonalization in regression*. Retrieved October 5, 2009, from http://www.creative-wisdom.com//computer/sas/collinear_deviation.html

Yu, C. H. (in press). A model must be wrong to be useful: The role of linear modeling and false assumptions in theoretical explanation. *Open Statistics and Probability Journal*.

Yu, C. H. & Shawn, S. (2003). Evaluating spatial- and temporal-oriented multi-dimensional visualization techniques. *Practical Assessment, Research & Evaluation, 8*(17). Retrieved September 30, 2009 from http://PAREonline.net/getvn.asp?v=8&n=17

Zhang, T., Ramakrishnon, R., & Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. *Proceedings of the ACM SIGMOD Conference on Management of Data*, 103-114