Vendlinski, T., & Stevens, R. (2000). The Use of Artificial Neural Nets (ANN) to Help Evaluate Student Problem Solving Strategies. In B. Fishman & S. O'Connor-Divelbiss (Eds.), *Fourth International Conference of the Learning Sciences* (pp. 108-114). Mahwah, NJ: Erlbaum.

The Use of Artificial Neural Nets (ANN) to Help Evaluate Student Problem Solving Strategies

Terry Vendlinski, Ron Stevens
UCLA IMMEX Project, 5601 W. Slauson Avenue #255, Cluver City, CA 90230
Tel: 310-649-6568, Fax: 310-649-6591
Email: vendlins@mit.edu, immex_ron@hotmail.com

Abstract: This paper describes a technique for developing and analyzing detailed models of complex student problem solving, and methods to measure the reliability and validity of these models. Specifically, we use the Interactive Multi-media Exercises (IMMEX) system to record the specific steps students use to solve open-ended problems. While IMMEX has been used in numerous academic disciplines, the research documented in this paper relies on biology and chemistry students. We analyze thousands of such performances using artificial neural networks as a data-clustering tool that aggregates student performances without a priori knowledge of those performances and without the limitations imposed by comparing these performances to "experts." The resulting clusters serve as a rich source of assessment information, and can provide students and educators with the meaningful practical feedback necessary to improve learning. Finally, we analyze these clusters and explore the data features that influence the reliability and usefulness of such a tool.

Keywords: assessment, science education, artificial intelligence, meta-cognition

Introduction

Although the human mind has the capacity to recognize complex patterns and we humans use this ability routinely during our daily activities, these capabilities have limits. The five human senses have evolved to detect patterns or activity in certain limited ranges. An often unnoticed limitation is our capacity to perceive in only 3, 2 or 1 dimensions. While this latter limitation may seem to be of little importance in our daily encounter with the 3-dimensional world, it increasingly constrains our ability to describe (and ultimately to explain) the complex interaction of factors occurring in the world around us. The need to consider larger dimensionality, however, is not only important to the so called "hard sciences."

Increasingly, social sciences such as education seek to explain outcomes based on the simultaneous interactions of numerous variables. Representing more than three of these interactions in physical space, however, is virtually impossible without the aid of some mathematical model. Consider this, for example. A teacher notices that three students (Student 1, Student 2 and Student 3) have the test scores 92, 73 and 87, respectively. In this case, one could easily order the scores linearly (73, 87, and 92) in order to help assess what each outcome means. But the determination may be more difficult if another variable (say a test on a different concept within the same domain) is added.

If Students 1, 2, and 3 now have test score pairs (92, 66), (73, 85), and (87, 71), respectively, the meaning of the outcome may be less obvious. This is especially true if we want to compare student performance or understand the meaning of these outcomes. Statistical and other mathematical models are often developed in an effort to explain such multi-variable events. For example, the arithmetic mean (average) is often used to aggregate these multiple measures into a single descriptive outcome. If, however, the input variables (scores in this case) measure widely disparate attributes, such an average may not be very descriptive. In addition, we lose important information in the process. Such is the case in this example. In each case the students have an average score of 79. Knowing only this average score tells us nothing about the trends in student performance nor can it suggest changes to improve the performances that generated such scores. In this instance, a different mathematical construct may be more enlightening. Graphing the scores on a coordinate plane allows us to simplify complex patterns and can provide new insight into relationships between data.

In this case, each of the two scores a student received could be represented along an axis (the horizontal axis represents score on test 1; the vertical axis the score on test 2). Since we are interested in two scores, the graph has two dimensions. A third score could easily be represented with a third dimension. By using this graphical representation of the numerical scores we have not only preserved all the information in the original data set (unlike the average statistic computed above), but might also begin to develop additional hypotheses from our data. For example, the close proximity of Students 1 and 3 suggest they have similar performances on the two tests, while Student 2 achieved an identical average with very different performances. It should be noted that while the example here uses student performances on tests, the technique could be used to group (or cluster) other quantifiable performance measures as well. This paper will develop the use of this technique using a detailed, multi-dimensional performance metric which not only allows educators to determine how a student answered, but also how the student *arrived* at an answer. Unfortunately, graphical representations intended to cluster performances that involve more than 3 dimensions are difficult to visualize, and various statistical methods often remedy the problem by sacrificing some information contained in the data variables (as the use of the average above demonstrated). Consequently, we have studied and applied other techniques, especially those developed in the computer science field of Artificial Intelligence, to accurately cluster similar multi-variant performances. In particular, Kohonen artificial neural networks have demonstrated an ability to accurately cluster performances described by multiple descriptors and without the need for a priori knowledge of the data to be clustered. Using this tool, we have previously derived novel models of complex problem solving in education levels spanning high school through medical diagnosis (Stevens, et al., 1999). While the exploration and derivation of new models of performance is of fundamental importance, it is also important to develop measures that indicate the reliability and validity of these models and the factors that affect these psychometrics. This paper briefly describes these networks and how they were used to cluster thousands of high school student problem solving performances from the Interactive Multimedia Exercises (IMMEX) system. We then explore the features of the data that influence the reliability of these classifications, including the use of supervised ANN.

Like other ANN trained with unsupervised learning, Kohonen nets "find" clusters of data points in a large data set by moving a pre-determined number of "markers" to the mathematical center of the data. The neural net locates this center by moving the markers until the distance between them and the data points in the data cluster are minimized. Unlike supervised learning, the learning of unsupervised networks proceeds without comparing the output of the network to some external measure of what "should be." Instead unsupervised ANN are designed to minimize a built-in distance metric. We designed the unsupervised ANN used in the initial stages of the research to minimize Euclidean distance between the "markers" and data points. The supervised ANN employed in the later stages of the research described in this paper were effectively employed to refine the data clusters suggested by unsupervised learning and they improved the reliability of the final data clusters dramatically.

The IMMEX System and the Data it Produces

Traditional forms of pencil and paper assessment often lack the ability to accurately record how a student solves a problem from its presentation to its solution. While this is especially true for multiple choice type tests, the transitions students make in solving open-ended problems can also be unclear. Many designers of computer assessment systems have attempted to improve student assessment by giving their software "intelligence" in the form of templates that "experts" would use to solve the problem. Primary criticisms of these "expert" systems include their inability to measure novice problem solving, the difficulty discerning exactly how experts actually solve a particular problem, and the limitations these systems place on what is considered an "acceptable" student solution strategy (Lesh & Kelly, 1996). In each case, the expert system imposes severe, pre-determined limitations on how a student may solve a problem. The Interactive Multi-media Exercises (IMMEX) suite of software seeks to minimize many of these obstacles by presenting a problem, then allowing the student to select the path (s)he feels is best to reach a solution. Unlike other approaches (e.g. Mislevy, et al., 1999; Radlinski & Atwood, 1998), we specifically avoided developing a student model a priori or developing a static "expert" model of an acceptable student solution. Rather, we sought only to cluster "similar" student performances based on what the performance data itself contained. As students generate new performances, we can train fresh ANN to include these new observations. In this way the clusters are "adaptive" in nature. As such we seek to develop and refine a tool to inform assessment, not a substitute for human assessment.

In IMMEX, the path to a solution is comprised of a series of steps (often between 15 and 90 steps). Each step involves a student choosing one of a number of menu items from pull down menus in a standard graphical user

interface. As the student investigates the information provided by a selected menu item, IMMEX records this selection in a database for future analysis. This process of selecting menu items continues until the student has correctly solved the problem or has exhausted a predetermined number of attempts at a solution. When analyzed, a student's solution to a particular problem not only details how the student answered the problem, but also the information (and in what order they obtained the information) required to arrive at their answer. Although the available information is limited to predetermined menu items, unlike "expert" systems the path through the information is not predetermined. In addition, the information available in menu items is robust and is intended to provide more than enough information to solve the problem in multiple ways (Hurst, et. al, 1996). The introduction of multiple pathways to a solution, however, can create a hardship in assessment (both summative and formative). On the one hand, the lack of a predefined expert template allows students the flexibility to demonstrate the strategy that makes sense to them. On the other hand, it is difficult to separate the most effective student strategies from merely acceptable strategies without predefining those strategies or studying hundreds of different student solutions. In addition, as the number of menu items increases, the possible paths to a solution can increase exponentially. Obviously, this parallels the dimensionality problem discussed in the opening paragraphs of this paper. Training Kohonen ANN to cluster problem-solving strategies has proven an excellent mechanism to inform assessment and to provide students and educators with the meaningful and practical feedback necessary to improve learning.

How IMMEX Represents Student Performance

IMMEX represents student performance both graphically and numerically. The graphical representation presents each menu item as a distinct, labeled rectangle. As a student moves from one menu item to the next, a new rectangle is added to the graphic along with a line indicating the order in which the menu items were selected. The graphical representation of a hypothetical student performance is shown is Figure 1 below.

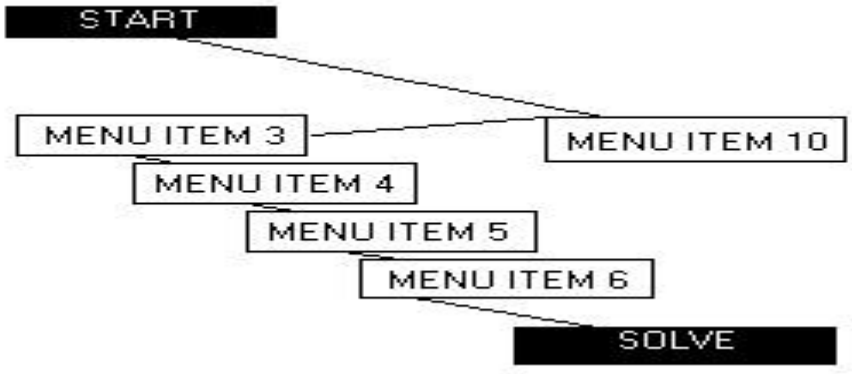

Figure 1. IMMEX graphical representation of a hypothetical student performance

In addition to this graphical representation, IMMEX provides a numerical representation of the same data. It is this numerical representation that is ultimately presented to the neural networks for analysis. The numerical representation of the hypothetical student performance from Figure 1 is shown below:

Student	1	2	3	4	5	6	7	8	9	10
001	10	0	4	5	6	11	0	0	0	3

Figure 2. IMMEX numerical representation of student performance

Figure 2 shows the performance of Student 001. This representation details, in columnar form, how the student proceeded through an IMMEX problem. After the initial presentation of the problem (called "START" in the graphical presentation and considered to be menu item 1 in the numerical representation), the student proceeded to menu item 10. Therefore, the first column shows the student moving from menu item 1 to menu item 10. The student next moved from menu item 10 to menu item 3. This transition is shown in the far right column of Figure 2. The student moved from item 3 to 4 (the third column of numbers) and so on, until (s)he solved the problem ("SOLVE" in the graphical presentation and menu item 11 in this numerical representation). The numerical

representation, therefore, shows menu item 11 below menu item 6 in Figure 2 and 0 for all remaining, but unselected, menu items. In IMMEX these menu items each represent specific information about the problem at hand. For example, a menu item might provide the results of a flame test in an analytical Chemistry problem or the results of a blood test in a genetics problem. The bottom row of numbers in Figure 2 forms a single, multi-dimensional data vector presented to the Kohonen neural network. Each individual student performance vector is termed an *exemplar* in the discussion below.

Training Kohonen ANN to Cluster IMMEX data

The designer of Kohonen ANN must set a number of parameters when the network is created. Because there are virtually no explicit rules for determining the value of such parameters, experience and testing are useful in choosing optimal values in such designs (Principe, et. al, 1997).

Trial and error suggested the value of most network parameters in this research. Because both the clusters of data and an artificial neural network's ability to find these clusters are easiest to visualize in two dimensions, we initially developed networks that would routinely find eight clusters of two-dimensional data. Ultimately we constructed a Kohonen ANN with parameters set to optimally cluster this data.

The same neural net developed for two-dimensional data was then used to determine the data clusters in two multi-dimensional IMMEX data sets ("True Roots" and "Desperately Seeking Solution"). In each case the only network parameter we changed for different data sets was the number of input dimensions. This change was necessitated by the nature of each specific problem. For example, in the analytical Chemistry problem, "Desperately Seeking Solution", each student could choose up to 90 different menu items. As described above, this is akin to comparing student performances across 90 dimensions simultaneously. Therefore 90 inputs are required. The "True Roots" genetics problem requires less than 60 dimensions.

Nets Trained with "True Roots" Data

The "True Roots" IMMEX problem involves students in the application of genetics concepts and problem solving skills to discern the true biological parents of a fictional high school student, Leucine. The students can select from 53 menu items (plus the start item) in their search for Leucine's actual biological parents. A Kohonen ANN was trained on two separate occasions with 642 solved performances. The ANN grouped these performances into 25 data clusters. Subsequently 1571 solved and unsolved performances (including the original data set of 642 solved performances) were fed into each of the two nets. The clusters produced by each net were then checked for inter-net reliability. The nets disagreed on how to cluster only 6 (less than 0.4%) of the almost 1600 performances ($\chi^2 = 0$, $p \approx 1$, $df=24$).

Inevitably, the question arises of what constitutes an appropriate data set for training an artificial neural network. Ultimately, the answer is that one should use a set of data that contains exemplars truly representative of the performances one eventually desires to cluster. Included in such deliberations must also be the eventual use of the clustering. These questions are similar to the ones that should be asked of any assessment. In training ANN, however, we must ponder the question of whether training with only successful student performances is varied enough to cluster both the successful and unsuccessful student attempts to solve the problem. The decision to use successful student solution strategies to train the net in this research was made for two reasons. First, experience suggests that students successfully solve IMMEX problems using a variety of strategies and that these strategies vary over time. In fact, the "True Roots" data contains successful performances in which the student obviously guessed at a correct solution. In general, adding unsolved performances to the training set would not diversify the solution strategies on which to base the data clusters discovered by the net. In fact, adding unsuccessful performances to the training set would merely add a lot of "noise" (disparate, infrequently used, and unsuccessful strategies) to the data. This in turn could require the net to dedicate markers to inappropriate strategies used by small numbers of students, while simultaneously forcing the net to group successful student strategies into larger clusters of more generalized strategies. Secondly, using a wide variety of successful performances ensures that unsuccessful strategies are ultimately compared to successful ones. In this way, we are able to determine how an unsuccessful strategy correlates with a successful one, and how a student might change the former to achieve a more successful outcome. In many ways, this resembles the use of "expert" templates while avoiding many of their pitfalls. In addition, unlike "expert" performances, these successful performances describe a wide range of strategies, which allows the ANN to classify successive student problem solving attempts as these attempts move

from novice-like to expert-like performances over time. Such longitudinal movement is evident in the "True Roots" data (Stevens, et. al, 1999).

Nets Trained with "Desperately Seeking Solution" Data

The IMMEX problem titled "Desperately Seeking Solution" presents students the problem of identifying an unknown compound and provides various analytical (generally physical and chemical) properties students can use to identify the compound. The problem allows students 89 possible tests in their attempt to solve this problem. We trained nets with 402 solved performances and used these nets to cluster 949 completed (solved and unsolved) performances. Nevertheless, the clusters produced by these nets revealed little information about the different strategies students had used to solve various sub-problems. We, therefore, retrained the Kohonen ANN with a different form of data. Instead of representing the order a student visited a specific menu item (described above), we replaced all non-zero item numbers in the data with the value "1". In effect, this replaces the order a student visited specific menu items with the much less detailed fact that (s)he merely visited the item sometime during his or her attempt at solving the problem.

The results of training Kohonen ANN with data representing if (not the order) a student visited a menu item in this problem produced performance clusters that were more uniform, easier to discern, and ultimately more useful for analysis of student performances. Researchers felt the clusters produced by retraining the ANN in this way better identified the specific strategies the student used to solve various sub-problems of the overall problem presented. In addition, the researchers generated 8 mock performances for the nodes they felt to be most important. Using the new nets, the researchers accurately mapped a specific performance to a particular node 100% of the time. A major drawback of training the ANN with unitary data, however, was the lack of consistency between different network training sessions. In this case, the same net trained with identical data on three separate occasions grouped approximately 60% of the performances it considered similar in the first round of training with those same performances in later rounds. However, when one accounts for the sub-groups that moved en masse from a single node in the first data clustering to a secondary node in the second and third data clusters, the consistency between nets increases to more than 80%. Furthermore, analysis of the data suggests no significant difference between the clusters produced by the final two nets ($\chi^2 = 17.2$, $p>.83$, $df=24$). We believe the net is discerning sub-groups of student performances that roughly correspond to one or two other nodes, but is having difficulty classifying these performances with certainty. While we have yet to investigate how these primary and secondary nodes are conceptually related to one another, we have demonstrated how we can use supervised training to further refine our initial clusters and increase the reliability of these clusters to approximately 90%.

Supervised Training

Based on the statistical analysis of the clusters when controlling for sub-groups, we developed supervised networks to further refine the clusters discerned by unsupervised learning. Initially we identified the student exemplars in the "Desperately Seeking Solution" data that remained clustered in the three training sessions described previously. We then used these exemplars (both solved and unsolved performances) to train a supervised ANN. Unlike unsupervised training which uses an internal distance metric to determine similarity between performances, supervised training presents the ANN an output the user expects from a specific input. In this case we trained the ANN by providing it an exemplar and the node at which that exemplar should be clustered. When we subsequently clustered student performance data using these trained supervised networks, more than 87% of the performances clustered at the same node on four separate occasions ($\chi^2 = 6.55$, $p\approx1$, $df=23$) (see Note 1). Furthermore, these clusters revealed discrete, often subtle, variations in student performance.

As with unsupervised learning, the supervised ANN divided 949 student performances from the "Desperately Seeking Solution" problem into 25 clusters. These clusters grouped performances ranging from students merely checking the overt properties of a specific item then guessing at that unknown's identity to performances in which students ran a series of detailed tests, then attempted to identify the unknown compound. Of particular interest was the ability of this ANN to routinely tease apart large clusters of student performances into sub-clusters that represented significant refinements in problem solving technique. For example, one cluster identified 63 student performances which investigated all the properties of an unknown, then reacted the unknown with 13 different salts to determine the unknown's identity. A second cluster grouped the 41 performances that conducted a flame test on an unknown, then reacted that unknown with only two of the 13 salts to identify the unknown. Finally, a third cluster grouped the 37 students that performed the same tests as the second cluster and an

additional pH test on the unknown. While we made no attempt to discern which cluster was "better", one can easily argue that the third strategy may be more appropriate to identify an unknown acid or base, while the second strategy would quickly discover an unknown with a distinctive flame test and the Chloride (Cl^{-1}) ion.

To test that the patterns researchers discerned at each node were, in fact, the same as those "recognized" by the ANN, we generated 23 mock performances (two nodes generated by this ANN currently contain no performances). Using the supervised nets, the researchers accurately mapped a specific performance to a particular node 100% of the time. Moreover, by generating mock performances that contained only those menu items chosen by 60% or more of the students at a node, the ANN was able to accurately place that mock performance 92% of the time. In fact, the only mock performances not correctly identified by researchers using this 60% threshold were those that represented clusters of only two student performances.

Conclusion

As human beings strive to more fully understand and explain the world around them, the need to develop sophisticated models that allow for the simultaneous interaction of numerous variables becomes more pronounced. This is as true for the "hard" sciences as it is for the "social" sciences, including education. While two and three-dimensional models are easy to visually interpret, they do not have the degrees of freedom necessary to model complex interactions of events of interest in the world surround. Of particular interest in this paper is our ability to accurately model and assess student understanding. Assessments of academic achievement based solely on what a student answered can increasingly be replaced with information about why a student arrived at a certain answer or the thought processes they used to reach a certain conclusion. But these increasingly sophisticated tools require more advanced methods of analysis to make sense of the new information at our disposal. This paper generally discussed the IMMEX technology and the wealth of information it provides about student problem solving. In addition, it briefly discussed the difficulty of comparing performances described by large numbers of variables (multi-dimensions). The paper introduces the concept of the Kohonen Artificial Neural Network (ANN) as a useful data clustering tool that, in this case, can help us discern patterns in the data, which may be too complex or too subtle for humans to detect. These ANN can "find" such patterns without any a priori knowledge about the data.

The Kohonen ANN detailed in this paper was able to consistently cluster data describing student performances while solving genetics, and analytical chemistry problems. In the first problem type, the network consistently clustered more than 99% of almost 1600 performances and allowed researchers the opportunity to study quality of student performance, longitudinal student improvement, and variability between classrooms. In the second problem, the results suggested a practical way to assess the data clusters produced by an ANN. This method required researchers to generate "mock performances", present these performances to the trained ANN and validate they were, in fact, clustered around the anticipated node. Although the unsupervised nets produced from unitary data in the last problem set were difficult to replicate with high levels of consistency, evidence that groups (not individuals) move across nodes suggests this was attributable to the presence of sub-groups of student performances. While difficult to consistently associate with one or more of the larger clusters, the results suggest these sub-clusters are not caused by random chance, but are actually present in the data. In fact, when supervised ANN were trained with those performances that consistently clustered together, not only did reliability improve to approximately 90%, but the granularity of the clusters also increased dramatically. As indicated above, these clusters suggest important subtle differences in the student strategies to solve for different unknown properties.

Finally, and most importantly, the techniques and ANN developed in this paper give us an alternative to "expert" systems and a starting point to begin refining subsequent assessment tools. Specifically these tools facilitate assessments of student problem solving that require more than a static, a priori model of how a student should solve a problem.

Endnotes

(1) All four training sessions produced one cluster with no student performances (i.e. a "dead node").

References

Hurst, K., Casillas, A., & Stevens, R.H. (1997). *Exploring the dynamics of complex problem-solving with artificial neural network-based assessment systems*. CSE Technical Report 444. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). University of California, Los Angeles, CA.

Lesh, R. & Kelly, A.E. (1996). A constructivist model for redesigning AI tutors in mathematics. In J.M. Laborde (Ed.) *Intelligent learning environments: The case of geometry* (pp. 134 – 156). New York, NY: Grenoble:Springer

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *On the roles of task model variables in Assessment*. CSE Technical Report 500, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). University of California, Los Angeles, CA.

Principe, J. C., Euliano, N. R., Lefebvre, W. C. (2000). *Neural and adaptive systems: Fundamentals through simulations*. New York, NY: John Wiles & Sons, Inc.

Radlinski, E. Robert & Atwood, Michael E. (1998). Augmenting intelligent tutoring systems with intelligent tutors. In Charles P. Bloom & R. Bowen Loftin (Eds.) *Facilitating the development and use of interactive learning environments* (pp. 73-101). Mahwah, N.J.: Lawrence Erlbaum Associates.

Stevens, R., Ikeda, J., Casillas, A., Palacio-Cayetano, J., & Clyman, S. (1999). Artificial neural network-based performance assessments. *Computers in Human Behavior*, 15, 295 –313.