12-31-1994

Use of Data Development Analysis for Certain Case Based Expert System Applications

Marvin Troutt
Southern Illinois University

Arun Rai

USE OF DATA ENVELOPMENT ANALYSIS FOR CERTAIN CASE BASED EXPERT SYSTEM APPLICATIONS

M. D. Troutt
Arun Rai
Aimao Zhang
College of Business and Administration
Southern Illinois University

ABSTRACT

Data Envelopment Analysis is a technique for comparing efficiencies of productive units based on their respective input and output data. The method is known as an efficient frontier technique and was developed originally by Charnes, Cooper, and Rhodes. This paper shows how the technique may be useful in an entirely different context — namely that of case or example based expert systems. In this latter area, it is desired to make decisions, such as acceptance or rejection of credit risks, based on examples which have previously been decided by an expert. This paper shows how Data Envelopment Analysis (DEA) may be used to develop an acceptance boundary for use in case based expert systems. Acceptability of cases is identified with cases which lie on or above the efficient frontier in the DEA sense. The method requires convexity of the acceptable set to hold as its major condition. The method also assumes that the accepted cases are accurately classified by the expert with respect to Type II errors.

1. INTRODUCTION

Data Envelopment Analysis (DEA) was introduced by Charnes, Cooper, and Rhodes (1978) as a method for comparing efficiencies of Decision Making Units (DMUs) based on their input and output data. In this paper, we describe how DEA can be used in the apparently unrelated setting of case or example-based learning. The application we discuss is that of acceptance or rejection of cases, such as credit risks, based on a vector of variable values for each case. In this setting, DEA can be used to first find a piecewise linear *acceptance boundary* and second to provide an acceptance/rejection rule for automatically classifying future cases. The contribution of this paper is therefore a demonstration of how to use DEA for case-based learning systems.

We limit our discussion to situations in which all continuous variables have the property called *conditional monotonicity* by Cronan, Glorfeld and Perry (1991). An example in the credit risk situation would be gross family income. Higher values of this variable could only improve the credit risk, indicating a monotone increasing relationship to acceptability of the case. A counter-example might be age (without credit insurance). Other things being equal, a credit applicant who is very old or very young may cause doubts about his or her acceptability. Such variables might be called *conditionally midmodal*, suggesting optimal value(s) more central than extreme. While our method considers only continuous variables which are conditionally monotone, we suggest a promising approach for handling midmodal variables in a later section. In case categorical variables are also present, our procedure can be carried out, in principle, for each category. Also it is assumed that all continuous conditionally monotonic variables are recoded, if necessary, so that acceptability of a case increases with the value of the variable.

The second major assumption we make is that of the convexity of the set of acceptable cases. Specifically, suppose x^1 and x^2 are the vectors of two acceptable cases. Then the convexity assumption requires that all cases of the form $\lambda x^1 + (1-\lambda) x^2$, $\lambda \geq 0$, $1 - \lambda \geq 0$ must also be acceptable. Interestingly, this assumption often appears reasonable but is typically violated by rule induction systems based on simple binary rules. For example, consider a simple set of rules for two continuous variables x_1 and x_2 given by:

If x_1 ≥ 5 then accept.
Else if x_1 ≥ 3 and
if x_2 ≥ 4.2 then accept.
Else reject .

The following two cases, (5, 2) and (3, 5), are therefore both acceptable by these rules. However, the case midway between them, namely (4, 3.5) would clearly be rejected. Hence we are able to make the important conclusion that when convexity of the acceptable set holds, simple binary rule based systems will not be appropriate for expert systems use. Moreover, use of rule based systems of this type would be expected to create numerous misclassification opportunities.

Rule based expert systems can be built in two ways at present. First, a knowledge engineer could elicit the knowledge from the expert. A variety of knowledge elicitation techniques have been reported in the literature (Byrd, Cossick, and Zmud 1992). No mention has been made in the literature of techniques that would enable adhering to the convexity assumption or on the implications of overlooking the same during the knowledge elicitation exercise.

The second way is known as rule induction. Given the high costs of traditional knowledge engineering methods, increasingly rule induction techniques such as Quinlan's ID3 and RPA (see Cronan, Glorfeld and Perry 1991) are being advocated. None of these techniques in fact have the option of imposing the convexity constraint during rule generation.

Convexity of the acceptable set of cases has therefore not been raised previously as an issue in the expert systems and rule induction literature and deserves extended study. In the meantime, it appears necessary to interact with an expert, if possible, to determine the suitability of this assumption in a particular application domain.

The main problem to be addressed in this paper can now be stated precisely as follows for each category determined by the categorical variables: Given a data set consisting of actions (accept or reject) and measurement vectors x, of conditionally monotone variables, and the assumption of convexity of the acceptable set, then determine an acceptance-rejection method for subsequent cases.

The present problem is apparently distinct from the statistical treatment of the discrimination problem. In that setting (see, for example, Morrison 1976), it is assumed that two or more distinct populations are involved. Here the cases all come from one population and the *acceptability* of cases is at issue. Presumably an expert forms some subjective impression of the expected profitability (or probability of default) of a case and then accepts those for which this value is sufficiently large (or small). Linear Discriminant Analysis (LDA) has been applied frequently in the present problem area. It is interesting to note that LDA assumes a convex acceptance set, since all cases within a given halfspace define the acceptable set of cases. DEA allows a piecewise linear acceptance boundary and therefore includes LDA as a special case. It is thus reasonable to expect the performance of DEA to be superior to LDA.

Little attention has so far been given to what may be called the *presenting population* from which all the cases are instances.[1] Here we propose a method which does not depend on knowledge of this population. In a later section we discuss how such information may be used, if available. However it is necessary to assume that the sample of cases to be used is at least *representative* of this input population. While this concept is clear from the layman's perspective it is a difficult one to operationalize. For example, is it sufficient that the sample have a centroid and variance-covariance matrix close to those of the population? How close is sufficient? Due to the complexity of these questions, we leave it to the judgement of the analyst and expert to pass on the acceptability of this assumption. A rule of thumb for regression analysis is to require a sample size of at least ten times the number of independent variables. The DEA method below fits a piecewise linear function to the data and is therefore somewhat similar to regression. From these connections we therefore suggest to use a sample with at least that number of cases per category.

Rules induced with differing ratios of the accepted and rejected cases can lead to very different rule systems — both individual rules and the ordering within them. Among other considerations, there needs to be a close match between the density of the training data and the density of the test data in terms of ratio of accepted to rejected cases. We experienced this practically when dealing with data from a large consumer corporation.

For example, consider a system developed using one hundred accepted cases and twenty-five rejected cases. Consider two test sets: Set A with twenty-five accepted cases and one hundred rejected cases and Set B with one hundred accepted cases and twenty-five rejected cases. Clearly, the rule structure is designed for a set B type situation and not a set A type situation. Present research on induction techniques pays little attention to this issue. However, the principle is similar to the prior probability concept in discriminant analysis.

The next issue to be raised is what may be termed *selectivity* of the expert and refers to the fraction of cases accepted by the expert from among those considered. In some settings, organizational policy may require an expert to accept cases at a greater or lesser rate (e.g., temporary shortage of funds, special new customer promotions). If

the expert is not restricted, then it may be presumed that he or she is also expected to decide on the appropriate fraction of cases to accept. This has practical significance for any resulting expert system. It may later be desirable to adjust the selectivity to a given level without seeking a new expert with a specified new selectivity rate. The method proposed here will also address the selectivity adjustment feature.

The last assumption deals with what might be called the *Error Type Emphasis* of the expert. Let v(x) be a measure of value (or expected value)[2] to the organization for a case with attribute vector x. If v° is the lowest value for which a case should be accepted, then the expert may be considered as estimating both v(x) and v°. That is, the expert accepts cases for which $v(x) \geq v^\circ$. In particular, it is assumed that all accepted cases in fact do have the property that $v(x) \geq v^\circ$. This can be seen as equivalent to assuming that the expert makes no Type II errors in which a case is accepted which should have been rejected. In turn, it must be assumed that little control over Type I errors has been exerted by the expert. The rationale for this assumption is that, in a credit risk situation, the Type II cost is likely to be substantially greater than the Type I cost which is merely an opportunity cost for rejecting a profitable case.

A summary of the assumptions is as follows:

1. *Conditional monotonicity* of all variables.
2. *Convexity* of the acceptable set.
3. *Representivity* of the sample cases.
4. *Selectivity* is unrestricted.
5. The cases contain no Type II errors.

2. THE BASIC RATIO DEA MODEL

It is useful to first review the basic ratio DEA model in its original context. Then the connection to the present problem may be developed. In the original context (Charnes, Cooper, and Rhodes 1978), output data vectors y_{rj} and input vectors x_{ij} are given for j=1 to N firms or units generically referred to as Decision Making Units (DMUs). This model finds output multipliers u_r and input multipliers v_i according to the following set of j_o=1 to N linear fractional programming problems.

$$\max h_{j_o} = \frac{\sum_r u_r y_{rj_o}}{\sum_i v_i x_{ij_o}} \qquad j_o = 1, N \quad (2.1)$$

$$s.t. = \frac{\sum_r u_r y_{rj}}{\sum_i v_i x_{ij}} \leq 1, \text{ for all } j \qquad (2.2)$$

$$u_r, v_j \geq 0 \; ^3 \qquad (2.3)$$

Note that a different problem results for each DMU. Thus the optimal weights depend on j_o as well. However, apparently by tradition in the DEA literature, the notation does not reflect this dependence (as in $u^*_{ij_o}$). Also it may be noted that some normalization is needed to yield a unique solution. Namely, if (u_r^*, v_i^*) is an optimal solution, then so is (cu_r^*, cv_i^*), for $c > 0$ by cancellation in all of the ratios. When the problem is solved using the Charnes and Cooper (1962) transformation, that method provides such a normalization automatically, as well as reduction to a linear programming problem.

To see the geometric behavior of the model in a related special case, suppose there is only one output which is unity for all DMUs. In that case, clearly u* cannot be zero for any DMU and the model (2.1 - 2.3) is easily reduced to

$$\min \sum v_i x_{ij_o} \qquad (2.4)$$

$$s.t. \; \Sigma v_i x_{ij} \geq 1 \text{ for all } j \qquad (2.5)$$

$$v_i \geq 0 \qquad (2.6)$$

Thus, consider Figure 1. If j_o=1 and v^1 is used, then clearly all points lie on or above $\langle v^1, x\rangle = 1$ and only point 1 lies exactly on the line. Hence the minimum optimal objective value of unity (1) is achieved. Likewise this observation holds for point 3. However, for any such supporting tangent at point 1 or 3, point 2 will lie strictly above it, unable to achieve the minimum value of unity. More generally, it is easily seen that the model (2.4 - 2.6) identifies the extreme points of the convex hull of the x data which are nearest to the origin in the sense of values of $\langle v^*, x\rangle$.

3. ACCEPTABILITY AND DEA EFFICIENCY

Consider the following association between the efficiency concept in DEA and the acceptability of cases. Let all acceptable cases be assigned an output value of unity. Let x^A be an acceptable case. Then by the conditional monotonicity assumption, we have that case x is also acceptable if

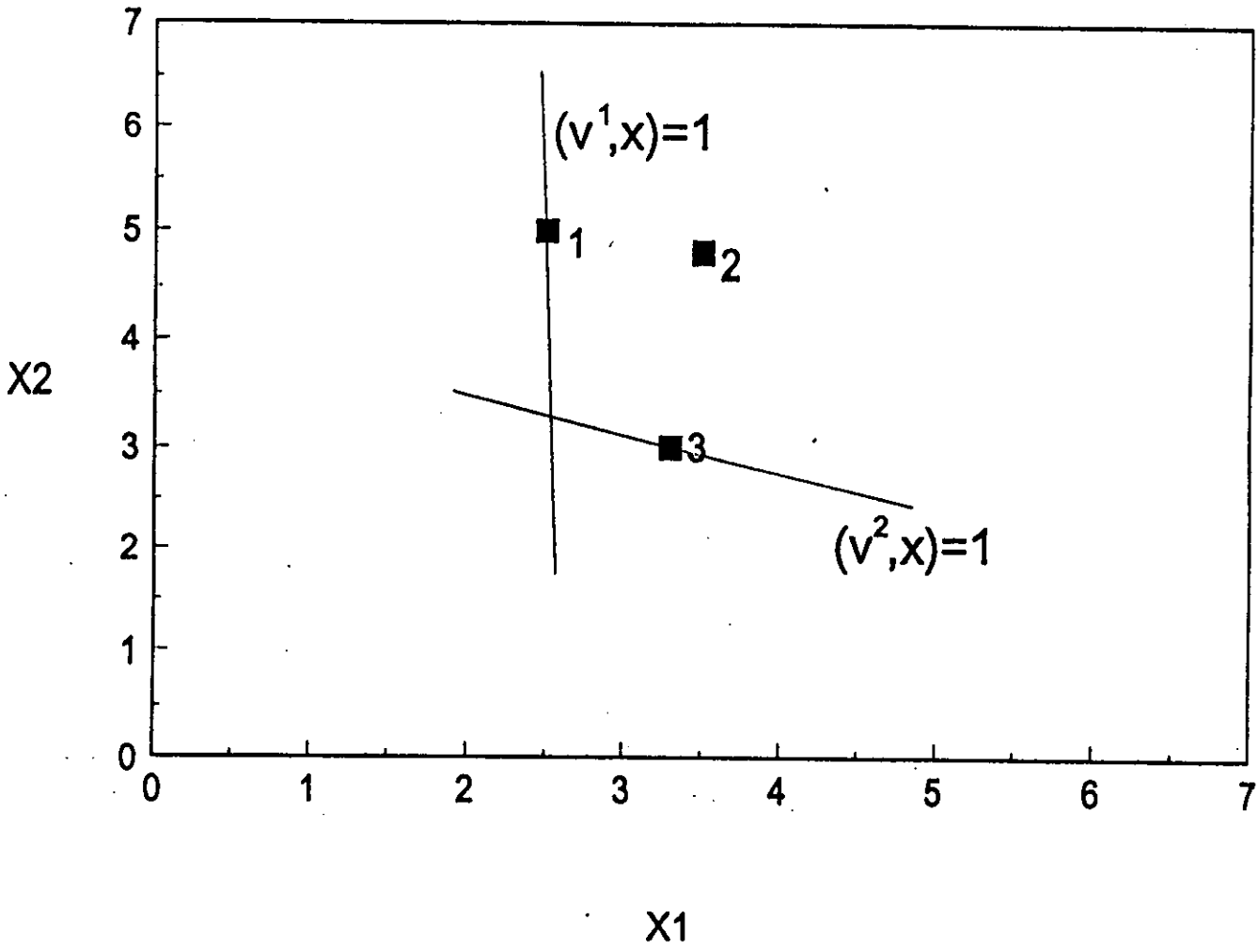

Figure 1

$$x_i^A \le x_i \text{, for all } i \quad . \tag{3.1}$$

Now consider any feasible DEA weights u_l and v_i. We have

$$\frac{u_1}{\sum_i v_i x_i^A} \ge \frac{u_1}{\sum_i v_i x_i} \tag{3.2}$$

since u_l , $v_i \ge 0$ and (3.1) must hold. It follows that point x is either of equal DEA efficiency to x^A or point x is technically *in*efficient. Hence the set of acceptable points consists of all points which are DEA efficient or technically inefficient, that is, all points on or above the efficiency frontier. Figure 2 shows the DEA efficiency frontier for a two variable set of hypothetical classified cases.

The proposed use of this efficiency/acceptability frontier is as follows. Given the set of training cases, compute the efficiency frontier for the accepted cases. That is, solve the DEA efficiency problems which identify the efficient accepted cases. Under the assumption that all these cases are truly acceptable to the organization, this frontier will be a conservative one. Namely, there may exist other acceptable cases not on or above the frontier which were either omitted by the expert in fear of Type II error, or which were never presented for review. Let E* be the subset of DEA efficient cases so identified. Then given a new case x^{new}, it is only necessary to determine whether x^{new} is on or above the frontier determined by the cases in E*. We now claim that the new case can be classified by the simple DEA efficiency program

$$\max \frac{u_1}{\Sigma v_i x_{ij}} \tag{3.3}$$

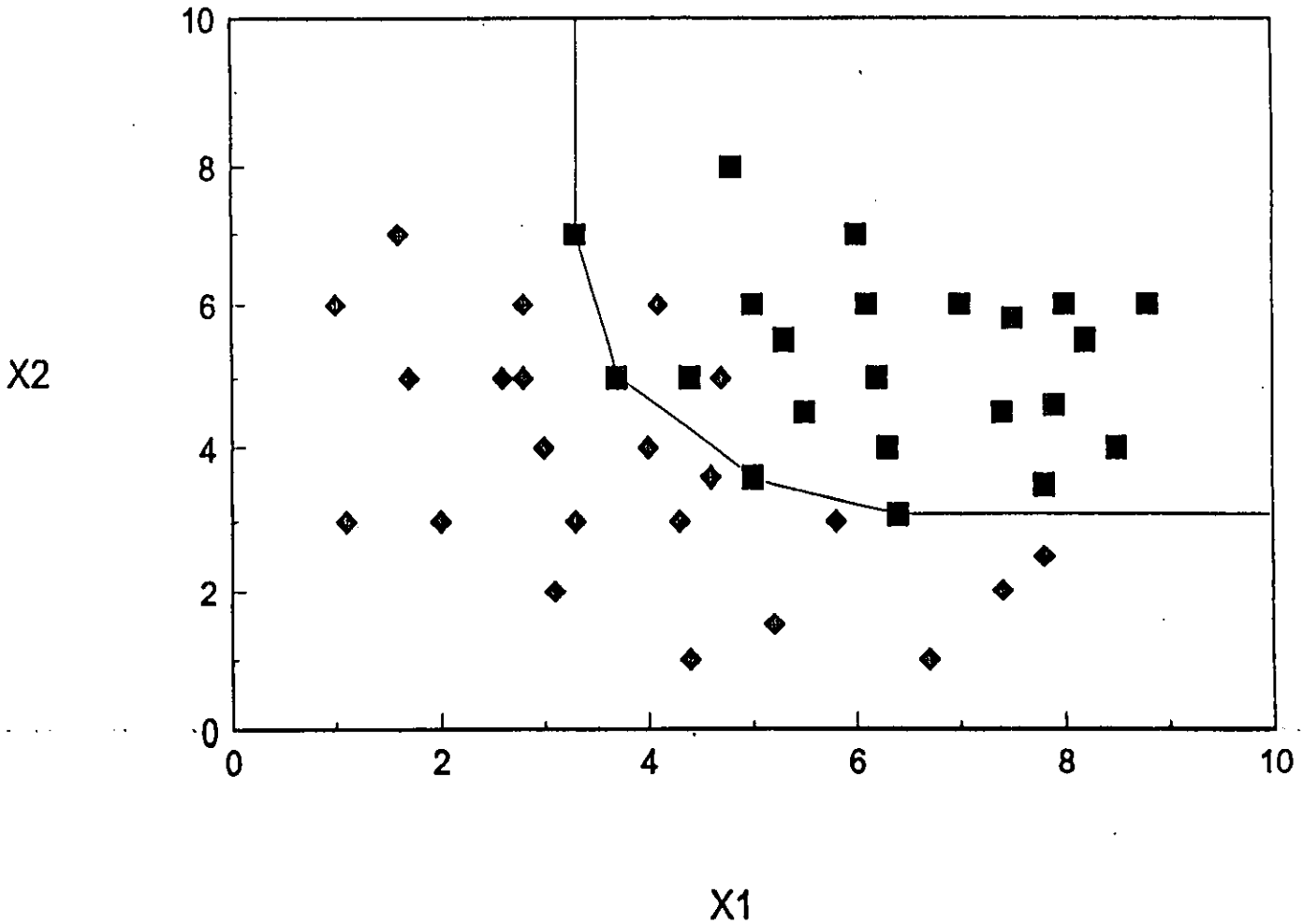

Figure 2

$$s.t.\ \frac{u_l}{\sum v_i x_{ij}} \leq 1 \text{ for all } x_j \in E^* \cup \{x^{new}\} \quad (3.4)$$

$$u_l\ ,\ v_i \geq 0 \quad , \quad (3.5)$$

along with the requirement that if x^{new} is DEA efficient then it lies within E*. This last condition is required in order to assure that a new point is a convex combination of existing efficient points. Without this restriction, a new point could become a new addition to E*. This would in effect change the contour associated with E*. The decision rule can therefore be stated as follows: Let u_l^*, v_i^* be optimal weights in problem (3.3) - (3.5).
Then

$$\textit{If } \frac{u_l^*}{\sum_i v_i x_i^{new}} < 1 \quad \textit{then accept the case} \quad . \quad (3.6)$$

Else if x^{new} is efficient and does not alter E^*
then accept the case . (3.7)
Else reject the case .

Condition (3.6) would indicate that the case is not DEA efficient and therefore lies above the frontier determined by the E*. Condition (3.7) determines whether the point lies on the frontier itself. It may be noted that condition (3.7) can be tested by determining the feasibility (Phase I problem) of the following linear program:

$$\max \quad L(\lambda) \tag{3.8}$$

$$\text{s.t.} \quad \sum_j \lambda_j x_{ij} = x_i^{new} \quad , \tag{3.9}$$

$$\text{for all i, j, } x_{ij} \in E^*$$

$$\sum_j \lambda_j = 1 \tag{3.10}$$

$$\lambda_j \geq 0 \quad , \tag{3.11}$$

where $L(\lambda)$ is an arbitrary linear function of the λ_j. All the above rules may be simplified to the program given in the following Theorem whose proof is provided in the appendix.

Theorem 1: A new case x^{new} is acceptable if and only if the following linear program is feasible.

$$\max \quad L(\lambda) \tag{3.12}$$

$$\text{s.t.} \quad \sum_j \lambda_j X_{ij} \leq x_i{*}\text{new}, \tag{3.13}$$

$$\textit{for all } i, j, x_{ij} \in E^*$$

$$\sum_j \lambda_j = 1 \tag{3.14}$$

$$\lambda_j \geq 0 \quad \text{for all j} \quad . \tag{3.15}$$

To summarize, the training or learning phase of our method uses DEA to find the efficient set of points, E*. Then the operational phase tests a given new case, x^{new}, using the program (3.12) - (3.15). If the program is feasible, the case is accepted and otherwise it is rejected.

4. DISCUSSION

The most critical assumption beyond convexity for use of the proposed method is that of the Type II accuracy of all accepted cases. When this assumption is valid (particularly for the E* cases), the system will be conservative in the following sense: Namely, assuming that the v(x) function is continuous, there are likely to be acceptable case occurrences which lie below the estimated frontier, which will later be rejected. On the other hand, if any of the E* cases should not have been accepted, then some future cases will be incorrectly accepted. Under our assumption that accepted cases are accurately classified, the resulting system will be expected to reject a few cases which might truly be acceptable but not conversely. We regard this as conservative, assuming that this type of error is less serious than accepting undesirable cases.

Given the above critical nature of the E* cases, additional attention to them is warranted. For example, when possible, it would be desirable to ask the expert to evaluate a random sample of cases along the frontier. A possible alternative is available when the selectivity rate can be reduced. Shifting the frontier upward by transformations of the form $E^*(d) = \{x + d : x \in E^*\}$ where d is a vector would evidently reduce the fraction of cases accepted while still using the same estimated frontier shape. At the same time, this would clearly increase the v(x) score of all the accepted points.

The foregoing discussion also suggests an approach to selectivity adjustment of the system at later times of operation. For example, consider E*(d) for vectors of the form d = ce where e is the vector of units. Using the training cases, or an updated version, as a sample distribution for the presenting population, values of c could be experimentally adjusted upward until a desired fraction of cases are accepted.

These results also point out the need for caution in interpreting misclassification rates in comparative studies of approaches to such systems. A typical procedure is to train the system (e.g. induce rules as in Chung and Silver [1992], Cronan, Glorfeld and Perry [1991], and Liang [1992]) and then compute misclassification rates on a hold out sample of the cases. When the DEA approach provides an accurate acceptance boundary, rules which are obtained from E* would be expected to give nearly perfect classification rates on a hold-out sample which lies well above the frontier. Namely, cases far above the frontier would not be expected to be affected by the nonconvexity of the rules near the acceptable frontier.[4]

In some systems, as noted earlier, it may be appropriate to assume that the expert has tended to achieve low Type I errors. In this case, it may be assumed that the rejected cases are more accurate. A possible modification can be suggested based on Figure 2 and the E*(d) translations above. Namely, let d be chosen such that the frontier is moved up just sufficiently to exclude all rejected cases. In

fact, it appears possible to choose d in such a way as to accomplish this exclusion while retaining a maximal number of accepted cases. It is left beyond the scope of this paper to provide a corresponding computational procedure. However such an upper and lower frontier may be determined, the possibility clearly exists for various averages to produce a single boundary for situations in which it is felt that Type I and Type II errors are more equal in impact.

5. CONCLUSIONS AND FURTHER RESEARCH

This paper shows how DEA may be used in simple acceptance-rejection systems when all variables are conditionally monotone and the true acceptable set is convex. If categorical variables are present, then the method should be carried out separately for each category. The method produces a conservative system when the expert data are assumed to be Type II accurate on all acceptance decisions.

A set of cases, E*, analogous to the DEA efficient subset is seen to play a critical role. When possible, the expert should recheck decisions in the vicinity of the corresponding efficient frontier.

Several questions arise for further research. First, if the analyst and expert believe that conditionally midmodal variables are present, some type of recoding needs to be considered for transforming them to conditionally monotone form. For example, if variable x has an optimum value of x^0, a transformation of the form $\tilde{x} = a - b(x-x^0)^2$, with a, b > 0, would yield a conditionally monotone replacement, $\tilde{x}$, for x.

Second, regression techniques might be applied to fit smooth contours to the E* cases. Such smooth contours, as for example associated with Cobb-Douglass production functions (see, for example, Charnes et al. 1988) might facilitate computation for selectivity adjustments.

Next it appears feasible to recompute E* from time to time due to learning by the expert. For example, a new case may be inserted into the training set if the expert has handled it outside the system. Also, if a case is later reconsidered and reclassified then its data could be changed. Following such changes, the E* set would need to be recomputed. In this way, the system can be dynamically adjusted.

Finally, it may be appropriate to assume convexity of the true rejection set for some applications — particularly those for which Type I error avoidance characterizes expedited expert behavior. It is conjectured that models similar to (3.12)-(3.15) may be similarly derived for situations such as these.

6. REFERENCES

Byrd, T. A.; Cossick, K. L.; and Zmud, R. W. "A Synthesis of Research on Requirements Analysis and Knowledge Acquisition Techniques." *MIS Quarterly*, Volume 16, Number 1, March 1992, pp. 117-138.

Charnes, A., and Cooper, W. W. "Programming with Linear Fractional Functionals." *Naval Research Logistics Quarterly*, Volume 9, Numbers 3 and 4, September-December, 1962, pp. 181-186.

Charnes, A.; Cooper, W. W.; and Rhodes, E. "Measuring the Efficiency of Decision Making Units." *European Journal of Operational Research*, Volume 2, 1978.

Charnes, A.; Cooper, W. W.; Seiford, L.; and Stutz, J. "Invariant Multiplicative Efficiency and Piecewise Cobb-Douglass Envelopments." *Operations Research Letters*, Volume 2, Number 3, 1988, p. 101.

Chung, H. M., and Silver, M. "Rule Based Expert Systems and Linear Models: An Empirical Comparison of Learning-by-Examples Methods." *Decision Sciences*, Volume 23, 1992, pp. 687-707.

Cronan, T. P.; Glorfeld, L. W.; and Perry, L. G. "Production System Development for Expert Systems Using a Recursive Partitioning Induction Approach: An Application to Mortgage, Commercial, and Consumer Lending." *Decision Sciences*, Volume 22, Number 4, September/October, 1991, pp. 812-845.

Hornik, D.; Stinchcombe, M.; and White, H. "Multi-Layer Feedforward Networks are Universal Approximators." *Neural Networks*, Volume 2, 1989, pp. 259-366.

Liang, T. "A Composite Approach to Inducing Knowledge for Expert Systems Design." *Management Science*, Volume 38, Number 1, January, 1992, pp. 1-17.

Morrison, D, F, *Multivariate Statistical Analysis*, Second Edition. New York: McGraw-Hill Book Company, 1976.

Tam, K. Y. "Neural Networks Models and the Prediction of Bankruptcy." *Omega*, Volume 19, Number 5, 1991, pp. 429-445.

White, H. "The Case for Conceptual and Operational Separation of Network Architectures and Learning Mechanisms." Discussion Paper 88-21. Department of Economics, University of California, San Diego, 1988.

Zahedi, F. "An Introduction to Neural Networks and a Comparison with Artificial Intelligence and Expert Systems." *Interfaces*, March-April, 1991, pp. 25-38.

7. ENDNOTES

1. This is similar to what White (1988) has called the *input space environment* in the neural network literature. (See also Hornik, Stinchcombe and White 1989.)

2. Note that ordinarily v(x) could be normalized to have range $0 \leq v(x) \leq 1$. Such a normalized v(x) could then be regarded as a membership function for the multivariate fuzzy set of acceptable cases. In the terminology of fuzzy sets, our method seeks a contour of minimum membership function value for acceptable cases.

3. It may sometimes be desirable to replace 0 by a small positive bound such as $\epsilon = 10^{-6}$. This is known as the non-Archimedean parameter. To see the value of this replacement suppose the data consists of a single output of unity for all DMUs and two inputs x_1 and x_2. Consider DMUs with input vectors (2, 3) and (3, 3). If the unrestricted $u_i^* = 0$ for this DMU, then these points would be declared of equal efficiency, despite the lower x_1 input of the first DMU. In the applications considered here, equal efficiency (acceptability) of these points is desirable so that $\epsilon > 0$ is not required.

4. Neural Network models have also been proposed for problems of the class being considered here. See for example Tam (1991) and Zahedi (1991). It is possible that training such a model on the E* set would be more time efficient, and perhaps more accurate, than on the whole data for similar reasons.

APPENDIX

Proof of Theorem 1: Let the acceptance criterion of feasibility of problem (3.12-3.15) be denoted as problem P1. To prove the "if" part of the theorem, suppose x^{new} is given and that λ^* is a feasible solution. Then $\Sigma\lambda_j^* x_{ij} \leq x_j^{new}$ for all j and $x_{ij} \in E^*$. Clearly x^{new} therefore lies above the E* frontier. For the "only if" part suppose first that (3.6) holds. Then it follows from (3.3-3.5) that x^{new} is not efficient and lies above the E* frontier. It follows that (3.13) must hold for λ so that P1 is feasible. Finally, if (3.7) holds, then x* must either coincide with a point e_i of E* (in which case $\lambda_i^* = 1$), or x* is a convex combination of a subset of E*. In either of these cases, the associated λ provides feasibility for P1. This concludes the proof.