A SEQUENTIAL INVERSE CLASSIFICATION PROBLEM

PARAG C. PENDHARKAR

Penn State Harrisburg
E-mail: pxp19@psu.edu

Abstract – This paper proposes a sequentialinverse classification problem (SICP) where a series of improvements are sought over time to classify a case from the currently classified undesirable class into another desirable class.These problems occur at personal and at organizational levels. Many insolvency, reengineering,and continuous quality improvement problems may be formulated as inverse sequential classification problems. Both linear and non-linear data envelopment analysis based procedures can be used to solve these SICP problems. This paper proposes and illustrates application of data envelopment analysis to solve SICP problems.

Index Terms - Inverse classification, Data Mining, Analytical Geometry, Linear Programming.

I. INTRODUCTION

Consider a situation where either a person or an organization is facing insolvency. The lending bank or shareholders would like the borrower to avoid this undesirable situation (going concern). Additionally, assume a competitive environment where lenders evaluate their potential borrowers periodically based on market situations that exist during that time period. In these situations of going concerns, a plan towards solvency needs to be designed. In designing such a plan, borrower input is sought where, due to the borrower's constraints, only some variables can be changed over the next time period to improve the financial situation of the borrower. It is likely that even these variables may change from one time period to next.Under these circumstances, traditional classification systems need modifications to aid financial planning.

We consider a binary classification situation where there are two classes—going concern and solvent. We label these classes as 0 (going concern) and 1 (solvent). For decision-making at time period t≥ 1, we assume that data from the previous time period (t-1) are available. This data from the previous time period reflects the current business environment where some of the borrowers are solvent and others delinquent/going concern. Let the vector $x_i=(x_{i1},\ldots,x_{im})^T$denote the current state of borrower i=1,…,n over m attributes. Let f: $(R^m)\rightarrow \{0,1\}$ be the classification function that is learned using the data for the time period (t-1) after managers subjectively label borrowers as belonging to class 0 or 1. The classification function may be linear or non-linear and once this classification function is learned, managers may use this function to work with the borrowers that are classified into class 0 and suggest a solvency plan while taking into account the variables each individual borrower wants to change over the next time period. Inverse classification problems were studied in literature before. For example, Mannino and Kaushik [1] proposed a cost-minimizing inverse classification problem and designed a genetic algorithm procedure to solve it. Pendharkar [2] proposed an inverse classification problem and showed that data envelopment analysis procedure can be used to solve it. The sequential inverse classification problem (SICP) in this paper is a slightly different problem. The SICP allows only certain user-selected decision-making attributes to be changed to solve the SICP while holding other attributes to a constant (existing) values. This selective approach to solving inverse classification problem provides managerial decision flexibility.

The rest of the paper is organized as follows. In Section 2, we propose a data envelopment analysis (DEA) based solution procedure to the SICP. In Section 3, we illustrate the application of the solution procedure to a problem, and in Section 4 we conclude our paper with a summary.

II. SOLVING SEQUENTIAL INVERSE CLASSIFICATION PROBLEM USING DATA ENVELOPMENT ANALYSIS

Data envelopment analysis is used for classification [3] and probabilistic classification problems [4] for over a decade. For SICP we assume that desirable class is labeled as 1 (solvency) and examples requiring changes currently are classified as belonging to class 0. Additionally, we make assumptions of monotonicity and representativeness. Monotonicity implies that higher values of all attributes increase the likelihood of an example being classified into class 0. Representativeness means that the sample under consideration is representative of the population and does not contain any particular biases that would make it different from the population of borrowers considered in decision time frame. Once these assumptions are satisfied, we can only focus on a set of borrowers that need improvements in their attribute values over the next time period. Let there are k<

nsuch borrowers that belong to class 0. The following steps are carried out during each decision time period.

Step 1: For a case under consideration, say $k_0 \in$ class 0, divide its elements into two sets X^Z and X^{NZ} where X^Z contains all the elements of the vector x_{k_0} that cannot be changed (i.e., zero improvements) and X^{NZ} contains all the elements of the vector x_{k_0} that can be changed over the next time period. The original elements of

$x_{k_0} = X^Z \cup X^{NZ}$.

Step 2:If k examples belonging to class 0, are labeled as j=1,…,k then estimate their optimal efficiency by solving following linear program for each j_0 =1,…,k:

Minimize ξ_{j_0},

Subject to:

$$\sum_j \lambda_j x_j^{NZ} - \xi_{j_0} x_{j_0}^{NZ} \leq 0,\ x_j^{NZ} \in X^{NZ},$$
$$\sum_j \lambda_j x_j^{Z} - x_{j_0}^{NZ} \leq 0\ ,\ x_j^{Z} \in X^{Z}, \qquad (1)$$
$$\sum_j \lambda_j = 1,$$
$$\lambda_j \geq 0, \text{ for all } j=1,\ldots,k.$$

Cases that achieve an efficiency score of $\xi_j^*=1$ are deemed efficient and create a frontier for class 0. If a case under consideration k_0 gets a score of 1 then the managerial recommendation would be to pick any one or more attributes from X^{NZ}andlower their values marginally by subtractingan infinitesimally small non-Archimedean amount δ. However, if $\xi_j^*<1$ then projections on the class 0 frontier are obtained before marginal reductions are applied. These projections p can be obtained for each element in the set X^{NZ} using the following formula:

$p^{NZ} = \sum_j \lambda_j^* x_{k_0}^{NZ}$, where λ_j^* are optimal values obtained by solving (1) where index $j_0 = k_0$.

A reader may note that multiple values of δ can be selected to obtain multiple solutions. This should not be surprising because higher values of δmean the case moves farther away from class 0 region and deeper into class 1. We now illustrate an application of this procedure in the next Section.

III. APPLICATION OF SICP USING HYPOTHETICAL AND REAL-WORLD DATASETS

We use a hypothetical dataset from Pendharkar [2] study to illustrate the application of the procedure. Table 1 illustrates this dataset where all examples belong to class 0.

Table 1: Hypothetical Dataset with Two Attributes

Case Number	x_1(Attribute 1)	x_2(Attrribute 2)
1	45	2.9
2	31	2.7
3	35	2.5
4	40	2.7
5	37	2.8
6	50	2.5
7	52	2.6
8	55	2.3
9	57	2.5
10	45	2.7

After applying the procedure, we get cases 2, 3, and 8 falling on the efficiency frontier with efficiency scores of 1. These cases define class 0 frontier. Since the dataset contains only two attributes, we can exhaustively solve the problem for all possible solutions. Assuming δ=0.01, these solutions are reported in Tables 2-4.

Table 2: Solutions with $X^Z=\phi$ and $X^{NZ}=\{x_1,x_2\}$

Case Number	x_1(Attribute 1)	x_2(Attrribute 2)
1	38.27	2.46
2	30.99	2.69
3	34.99	2.49
4	36.76	2.47
5	33.81	2.55
6	47.49	2.36
7	47.49	2.37
8	54.99	2.29
9	52.91	2.31
10	40.70	2.43

Table 3: Solutions with $X^Z=\{x_1\}$ and $X^{NZ}=\{x_2\}$

Case Number	x_1(Attribute 1)	x_2(Attrribute 2)
1	45	2.39
2	31	2.69
3	35	2.49
4	40	2.44
5	37	2.47
6	50	2.34
7	52	2.32
8	55	2.29
9	57	2.29
10	45	2.39

Table 4: Solutions with $X^Z=\{x_2\}$ and $X^{NZ}=\{x_1\}$

Case Number	x_1(Attribute 1)	x_2(Attrribute 2)
1	30.99	2.9
2	30.99	2.7
3	34.99	2.5
4	30.99	2.7
5	30.99	2.8
6	34.99	2.5
7	32.99	2.6
8	54.99	2.3
9	34.99	2.5
10	30.99	2.7

Individual case recommendations can be selected from the three tables. For example, if case 1 can only change thex_1 variable then it should aim to lower its current value of 45 to a new value of 30.99 (Table 4). Ifthe case 4 can change both variables then it should aim to change its current values to new values listed in Table 2. In cases with ratio variables where class 1 has higher means and class 0 has lower means, data can be transformed using reciprocals so that class 0 means are higher and model (1) can be used. To illustrate this situation, we use real-world bankruptcy prediction dataset from Pendharkar [2] study. This original dataset is illustrated in Table 5 for companies labeled as going concern. Table 6 illustrates the transformed dataset.

Table 5: Going Concern Class Financial Ratios

Obs.	Cash Flow/Total Debt	Net Income/Total Assets	Current Assets/Current Liabilities
1	0.0643	0.0156	1.0077
2	0.0351	0.0147	1.5046
3	0.0713	0.0205	1.3124
4	0.0109	0.0011	2.1495
5	0.1454	0.0500	1.8762
6	0.3703	0.1098	1.9941
7	0.0451	0.0263	1.6756
8	0.1227	0.1055	1.1434

Table 6: Transformed Going Concern Class Financial Ratios

Obs.	Total Debt/Total Cash Flow	Total Assets /Net Income	Current Liabilities /Current Assets
1	15.552	64.102	0.992
2	28.490	68.027	0.665
3	14.025	48.780	0.762
4	91.743	909.090	0.465
5	6.877	20.000	0.533
6	2.700	9.107	0.502
7	22.172	38.023	0.597
8	8.149	9.479	0.875

Let us assume that all companies aim to reduce their total debt and current liabilities to improve their financial situation over the next time period. Additionally, we assume δ=0.01 as before. We solve this problem first using the transformed dataset and get solutions shown in Table 7. Cases 4 and 6 get an efficiency score of 1.

Table 7: Solution using Transformed dataset

Obs.	Total Debt/Total Cash Flow	Total Assets /Net Income	Current Liabilities /Current Assets
1	7.82	64.102	0.48
2	8.51	68.027	0.48
3	6.62	48.780	0.49
4	91.73	909.090	0.46
5	3.76	20.000	0.49
6	2.69	9.107	0.49
7	5.55	38.023	0.49
8	2.72	9.479	0.49

Solution to original dataset then is reciprocal values of those shown in Table 7. These final solution values are shown in Table 8.

Table 8: Going Concern Class Financial Ratios

Obs.	Cash Flow/Total Debt	Net Income/Total Assets	Current Assets/Current Liabilities
1	0.128	0.0156	2.08
2	0.118	0.0147	2.08
3	0.151	0.0205	2.04
4	0.011	0.0011	2.17
5	0.266	0.0500	2.04
6	0.371	0.1098	2.04
7	0.180	0.0263	2.04
8	0.367	0.1055	2.04

CONCLUSION

This paper proposes a sequential inverse classification problem and suggests a DEA classification procedure to solve it.The SICP has multiple solutions and decision-maker has a choice to pick decision variables where improvement may be sought. In cases where cost information is available, decision-makers can easily incorporate such information in the proposed framework to suggest solutions that lower overall costs to make improvements.

The advantage of the DEA classification model is that classification boundary can be learned even when information about only one class is available. We made an assumption that examples are representative of actual cases for the time period. This assumptionis important as minor violations to the assumption may be tolerated as long as cases violating the assumption do not get an efficiency score of 1. We believe that the procedure will be generally robust for most real-world decision making situations.

REFERENCES

[1] M. V. Mannino, and M. V. Koushik, “The cost minimizing inverse classification problem: A genetic algorithm approach,” Decision Support Systems, vol. 29, pp. 283-300, 2000.

[2] P. C. Pendharkar, “A potential use of data envelopment analysis for the inverse classification problem,” Omega: An International Journal of Management Science, vol. 30, pp. 243-248, 2002.

[3] M.D. Troutt, A. Rai, and A. Zhang, “The potential use of DEA for credit applicant acceptance systems,” Computers and Operations Research, vol. 23, pp. 405-408, 1996.

[4] P.C. Pendharkar, “DEA models for probabilistic classification,” Computers and Industrial Engineering, vol. 119, pp. 181-192, 2018.

★★★