LACE and RFE: A combination of two algorithms applied to the forecasting of business profitability

Javier De Andrés
Pedro Lorca
Antonio Bahamonde
Juan José Del Coz
University of Oviedo

DIFFERENT APPROACHES FOR THE FORECASTING OF ECONOMIC PHENOMENA

- Regression analysis and classification techniques are two of the most popular approaches for the forecasting of economic phenomena
- Drawbacks:
  - Regression analysis: The functional form of the relationship among many variables is not easy to determine
  - Classification techniques: If the variable to analyze/forecast is continuous, it has to be discretized, and this implies loss of information

## AN ALTERNATIVE: THE APPROACH BASED ON PREFERENCES

- The objective is to predict if a certain company/individual/country will have in the future a higher level of a certain continuous variable than another company/individual/country
- Advantages for analysts and researchers:
  - The loss of information is smaller than using classification systems
  - There is no need to determine the exact functional form of the relationship among the variables under study

ALGORITHMS FOR LEARNING PREFERENCES

- The primary data set is a collection of preference judgments: pair of vectors (v,u) for one of which a preference is expressed
- In our problem, v and u are vectors describing features of two companies that are being compared
- The preference relation is determined by the variable under study, i.e. profitability

LACE (LEARNING TO ASSESS FROM COMPARISON EXAMPLES)

- This algorithm estimates a linear function f (the *assessment function* or *ranking function*) which maximizes the probability of having f(v)>f(u) whenever v is better than u
- This function is determined by a vector w
- The estimation of w is a NP-hard problem
- It can be approximated through an iterative procedure similar to that used in the induction of oblique decision trees
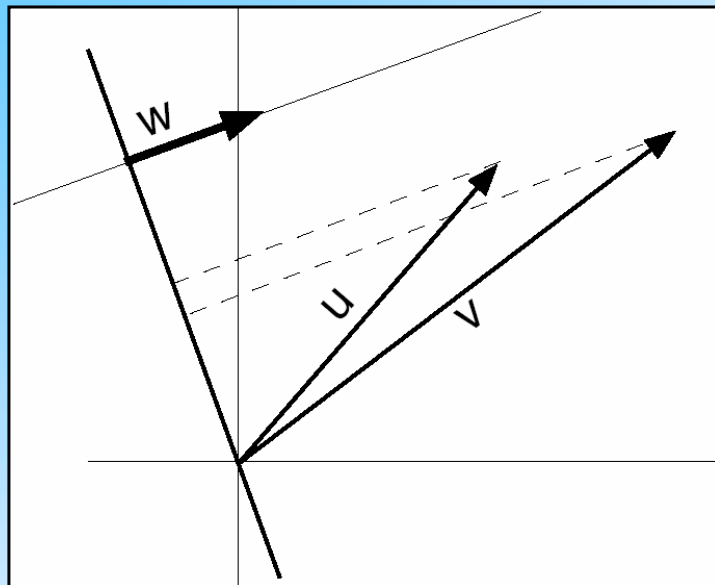
FEATURE SELECTION FOR LEARNING PREFERENCES

- Learners' generalization performances can be improved when they are given only the information supplied by relevant features
- We use the Recursive Feature Elimination (RFE), which is a backward feature elimination process
- Based in the results of LACE, a list of feature subsets is obtained, each one containing a different number of features
- Then, a metric-based method called ADJ (ADJusted distance) is used to choose the appropriate level of complexity required to fit the data, that is, the subset that leads to the better results

THE DATA BASE

- Medium sized companies: more than 100 employees and fewer than 250 (unlisted companies)
- Financial statements obtained from the Spanish edition of Bureau Van Dijk-Informa database
- Years under study 1998, 1999, 2000, 2001
- After the filters: 2,360 firms

PROFITABILITY

- In this study, profitabilty is the variable that determines which one is preferred from a pair of firms
- Prior research on small firms' profitability: neither with regression analysis nor using classiffication techniques good results are obtained
- Financial profitability ratio: quotient between the company's net profit and equity capital
- In order to eliminate distortions: average profitability for the years 2000 and 2001

THE PREDICTORS

- Average values for the years 1998 and 1999
- Limitation: Certain information not contained in the annual accounts
- Based on prior research eight dimensions were selected
- To avoid multicollinearity, each dimension is represented by only one financial ratio

Dimension	Variable	Code
Debt quality	$\dfrac{CurrentLiabilities}{TotalDebt}$	V1
Indebtedness	$\dfrac{EquityCapital}{TotalDebt}$	V2
Use of fixed capital	$\dfrac{TangibleFixedAssets + IntangibleFixedAssets}{TotalEmployment}$	V3
Debt cost	$\dfrac{FinancialExpenses}{TotalDebt}$	V4
Short-term liquidity	$\dfrac{CurrentAssets}{CurrentDebt}$	V5
Share of labour costs	$\dfrac{LabourCost}{AddedValue}$	V6
Size	Net Sales (EUR thousands)	V7
Average sales per employee	$\dfrac{NetSales (EURthousands)}{TotalEmployment}$	V8

DIFFERENT SETS OF VARIABLES

- 'Absolute' variables: labeled V1 to V8
- To avoid the sector effect an additional set of ratios ('relative' indicators): divide each ratio by the median for each branch of activity (labelled RV1 to RV8)
- To measure the dynamic aspects:
  - Variations of 'absolute' indicators (VARV1 to VARV8)
  - Variations of 'relative' indicators (VARRV1 to VARRV8)

THE TESTS

- Test n. 1: Each company is compared with the rest of firms included in the sample
- Test n. 2: Each company is compared only with the firms in the same NACE sector
- Test n. 3 : Each company is compared only with the firms with the same four digit code within the Spanish SIC

- Test n. 4:
  - Four groups of firms determined by the quartiles of the distribution of the financial profitability
  - Each company is compared with 10 randomly selected from the three groups not containing the company
- Test n.5: Similar to n. 4, the difference is that the two intermediate quartiles are deleted from the sample
- Test n. 6: The function inducted in test n. 5 is used to predict one to one comparisons between every pair of companies in the whole sample

THREE VERSIONS FOR EACH TEST

- Absolute: Only absolute variables (V1 to V8 and VARV1 to VARV8) are included in the set of predictors
- Relative: Only relative variables (V1 to V8 and VARV1 to VARV8) are included in the set of predictors
- Complete: All the variables are included in the prediction set

RESULTS OF THE TESTS FOR THE ABSOLUTE SET OF VARIABLES

Test	Error	Num. Var.	V1	V2	V3	V4	V5	V6	V7	V8	VAR V1	VAR V2	VAR V3	VAR V4	VAR V5	VAR V6	VAR V7	VAR V8
T1	30.84%	11	X	+	-	+	X	-	-	X	X	+	-	-	X	-	+	-
T2	30.02%	13	+	+	-	+	-	-	-	X	X	+	-	-	X	-	+	-
T3	35.73%	2	X	X	X	X	X	-	X	X	X	X	X	X	X	-	X	X
T4	27.82%	9	+	+	-	+	X	-	X	X	X	+	X	-	X	-	+	X
T5	17.99%	3	X	X	-	X	X	-	X	X	X	X	X	X	X	-	X	X
T6	29.60%	3	X	X	-	X	X	-	X	X	X	X	X	X	X	-	X	X

RESULTS OF THE TESTS FOR THE RELATIVE SET OF VARIABLES

Test	Error	Num. Var.	RV1	RV2	RV3	RV4	RV5	RV6	RV7	RV8	VAR RV1	VAR RV2	VAR RV3	VAR RV4	VAR RV5	VAR RV6	VAR RV7	VAR RV8
T1	31.49%	11	+	+	-	+	X	-	X	X	X	+	-	-	X	-	+	-
T2	30.69%	12	+	+	-	+	X	-	-	X	X	+	-	-	X	-	+	-
T3	29.50%	12	+	+	-	+	X	-	X	-	X	+	X	-	-	-	+	-
T4	27.21%	12	X	+	-	+	X	-	-	X	+	+	-	-	X	-	+	-
T5	15.88%	5	X	X	-	+	X	-	X	X	X	X	-	X	X	-	X	X
T6	28.97%	5	X	X	-	+	X	-	X	X	X	X	-	X	X	-	X	X

RESULTS OF THE TESTS FOR THE COMPLETE SET OF VARIABLES

Test	Error	Num. Var.	V1	V2	V3	V4	V5	V6	V7	V8	VAR V1	VAR V2	VAR V3	VAR V4	VAR V5	VAR V6	VAR V7	VAR V8
T1	29.96%	23	X	+	-	-	+	-	-	X	+	X	X	+	+	-	-	X
T2	31.70%	8	X	X	-	+	X	X	X	X	X	X	X	X	X	X	X	X
T3	35.76%	2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
T4	26.72%	13	X	X	-	+	X	-	+	X	X	X	X	X	X	-	+	X
T5	13.77%	14	X	+	-	+	X	-	+	X	X	X	-	-	X	-	X	+
T6	27.27%	14	X	+	-	+	X	-	+	X	X	X	-	-	X	-	X	+

Test	RV1	RV2	RV3	RV4	RV5	RV6	RV7	RV8	VAR RV1	VAR RV2	VAR RV3	VAR RV4	VAR RV5	VAR RV6	VAR RV7	VAR RV8
T1	X	+	-	+	X	-	+	X	-	+	-	-	X	-	+	-
T2	X	X	-	X	X	-	X	X	X	X	-	X	-	+	-	
T3	X	X	X	X	X	-	X	X	X	X	X	X	-	X	X	
T4	X	+	-	X	X	-	-	X	X	+	X	-	X	X	+	X
T5	X	X	-	X	X	-	-	X	+	+	X	X	X	X	X	X
T6	X	X	-	X	X	-	-	X	+	+	X	X	X	X	X	X

MAIN RESULTS

- Error rates are always lower than that of a naïve prediction system (T1, T2…T6)
- Error rates do not substantially drop when comparing only firms in the same sector (T2 and T3)
- Error rates drop only slightly when comparing firms from different profitability quartiles (T4)
- For the generation of efficient functions, only the exemplary cases are necessary (T5 and T6).

- The results of the tests are not substantially altered when using the 'relative' set of indicators
- When using the 'complete' set:
  - The error rates show only a slight decrement
  - There is a certain instability in the signs of the coefficients of the variables

DIRECTIONS FOR FUTURE RESEARCH

- The use of another techniques for the identification and deletion of irrelevant variables
- The study of listed firms, and the consideration of an extended set of possible features
- The study of other topics through LACE and RFE (i.e. financial distress prediction)