Discriminant Analysis *versus* Machine Learning Techniques. An Application to the Prediction of Insolvency in Spanish Non-life Insurance Companies

Authors:

Zuleyka Díaz Martínez [a]
José Fernández Menéndez [b]
Mª Jesús Segovia Vargas [a]

[a] Department of Financial Economics and Accounting I
[b] Department of Business Administration
Universidad Complutense de Madrid

Introduction

- Prediction of **insurance companies insolvency** has arised as an important problem in the field of financial research

- Most approaches applied in the past to prediction of failure in insurance companies are **traditional statistical techniques, such as Discriminant Analysis**, which use financial ratios as explicative variables. However, these variables do not usually satisfy statistical assumptions, what complicates the application of these methods.

Introduction

- A number of **non-parametric techniques** have been developed, most of them belonging to the field of Machine Learning, such as neural networks, which have been successfully applied to this kind of problems. However, their black-box character make them difficult to interpret.

- Other machine learning methods are more useful for economic analysis, because the models provided by them can be easily understood and interpreted by analysts.

Purpose of the paper

The purpose of this paper is to compare the predictive accuracy of three data analysis methodologies - a well-known parametric statistical technique (LDA) and two non-parametric machine learning techniques (See5 and Rough Set) - on a sample of Spanish insurance companies.

Structure of the paper

The paper is structured as follows:

- In first place, some concepts of the tested techniques are introduced.

- In second place, we describe the data and input variables.

- In third place, the results of the three approaches are presented, as well as the discussion and comparison of these results.

- Finally, we close the paper with some concluding remarks.

The See5 algorithm

Learning systems based on decision trees

↓

Different algorithms for automatic construction of decision trees ⟷ Different criteria followed to carry out the exhaustive and mutually exclusive partitions among the set of objects

- Statistics*: CART (Classification and Regression Trees)* (Breiman *et al.*, 1984)

- Machine Learning: ID3, C4.5, See5 (Quinlan, 1997)

The See5 algorithm

The criterion employed in See5 algorithm to carry out the partitions is based on some concepts from Information Theory:

- *Entropy* of a random variable *x*:

$$H(x) = \sum_x p(x) \log_2 \frac{1}{p(x)}$$

- *Conditional entropy* of *x* given *y*:

$$H(x/y) = \sum_{x,y} p(x,y) \log_2 \frac{1}{p(x/y)}$$

The See5 algorithm

Naturally, $H(x/y) \leq H(x)$

This reduction in the uncertainty is called:

- ***Mutual information*** between *x* and *y*:

$$I(x;y) = H(x) - H(x/y)$$

In a first time, Quinlan choose to make each partition the y_i-variable that provided the maximum information about *x* -variable, that is, he maximized

$I(x;y_i) \Rightarrow$ *gain*

The See5 algorithm

Because this procedure introduces a bias in favour of y_i-variables with many outcomes, the subsequent releases of the algorithm chooses the y_i-variable that maximizes the relation:

$$\frac{I(x\,;y_i)}{H(y_i)} \Rightarrow \textit{gain ratio}$$

Additionally, in order to avoid that an attribute could be only chosen because it has a low value for entropy, what would increase the *gain ratio*, the numerator of this relation should be big enough.

The See5 algorithm

A common problem for the majority of rules and tree induction systems is that the generated models can be quite adapted to the training set and, consequently, they will be very specific. This problem is known as ***overfitting***.

The most frequent way of limiting this problem in the context of decision trees consists on eliminating some conditions of the branches of the tree, in order to achieve more general models. This procedure can be considered as a ***pruning*** process.

The See5 algorithm

See5 incorporates a ***post-pruning*** method for an original fitted tree that consists in replacing a branch of the tree by a leaf, conditional on a predicted error rate:

- Suppose that there is a leaf that covers *N* objects and misclassifies *E* of them.

- This could be considered as a binomial distribution in which the experiment is repeated *N* times obtaining *E* errors.

The See5 algorithm

- From this issue, the probability of error p_e is estimated, and it will be taken as the aforementioned predicted error rate.

- A confidence interval for the probability of error of the binomial distribution is estimated.

- The upper limit of this interval will be p_e (this is a pessimistic estimate).

The See5 algorithm

- In the case of a leaf that covers N objects, the number of predicted errors will be $N \cdot P_e$

- If we consider a branch instead of a leaf, the number of predicted errors associated with a branch will be just the sum of the predicted errors for its leaves.

- A branch will be replaced by a leaf when the number of predicted errors for the last one is lower than the one for the branch.

Rough Set

•Rough Set (RS) Theory was introduced by Pawlak (1982)

•RS is a method for classificating objects

•Every object is characterized by some information and belongs to some class

•We use the information about the object to determine what class the object belongs to

Rough Set

- The information about the object consists on a set of values of some attributes
- It is represented in the form of an information table
- Every object is represented by a row in the table and every attribute by a column
- An additional column contains the class of the object.

Rough Set

- Information Table

object	Attrib.1	...	Attrib.k	class
X				Y_1
Y				Y_1
Z				Y_2
T				Y_2

Rough Set

- U: the set of objects

$$U = \{x, y, z, t, ...\}$$

- P: the set of attributes

$$P = \{att1, att2, ...\}$$

- Y: the set of classes

$$Y = \{Y_1, Y_2, Y_3, ...\}$$

Rough Set

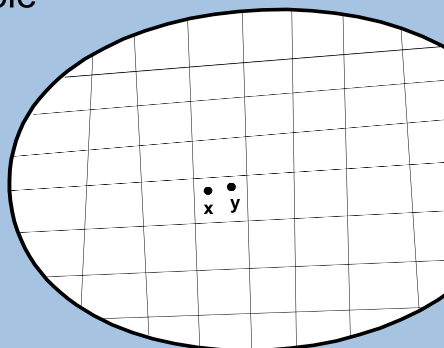
- Objects characterized by the same information are indiscernible in view of the available information
- We can so define an equivalence relation on the set U of the objects to be classified:

$$\forall x, y \in U \qquad xRy \Leftrightarrow att_i(x) = att_i(y) \quad \forall i$$

(i.e. its attributes have the same values)

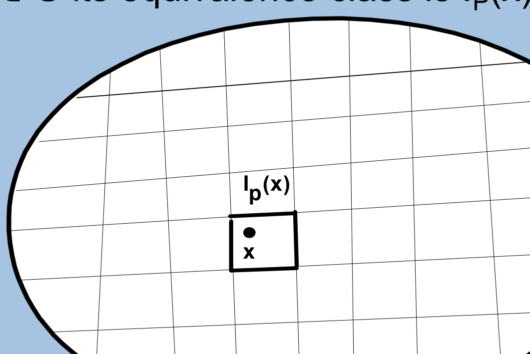
Rough Set

- This relation divides U in a set of equivalence classes. Objects in the same equivalence class are indiscernible



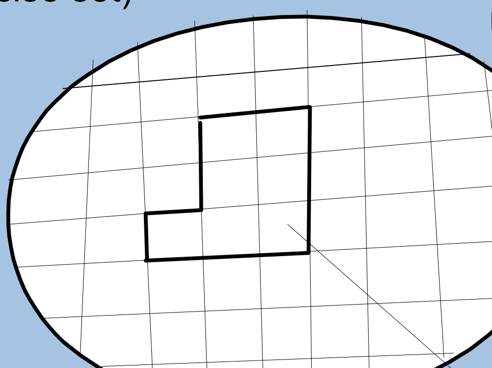
Rough Set

- We denote by I_P the set of equivalence classes generated by the set of attributes P
- If $x \in$ **U** its equivalence class is $I_P(x)$



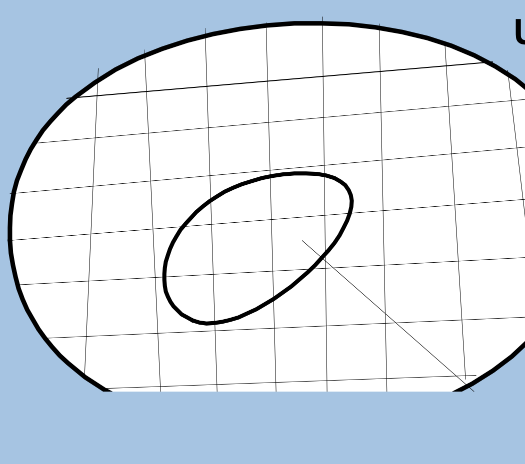
Rough Set

- Every equivalence class is an “**elementary set**”
- The union of elementary sets is a “**crisp set**” (a precise set)



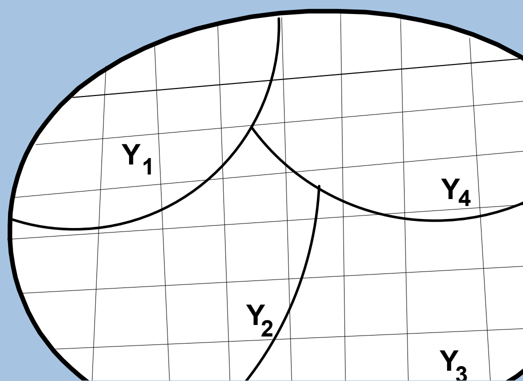
Rough Set

- Any set of objects that is not a crisp set is a rough set



Rough Set

- Every object $x \in U$ belongs to a class Y_i
- Every class Y_i is in general a rough set



Rough Set

- We approximate every rough set by means of crisp sets
- Every $Y_i \subseteq U$ is approximated by two crisp sets: the lower approximation $\underline{PY}_i$ and the upper approximation $\overline{PY}_i$
- $\underline{PY}_i$ is the biggest crisp set contained in Y_i
- $\overline{PY}_i$ is the smallest crisp set that contains Y_i

$$\underline{PY}_i \subseteq \mathbf{Y_i} \subseteq \overline{PY}_i$$

Rough Set

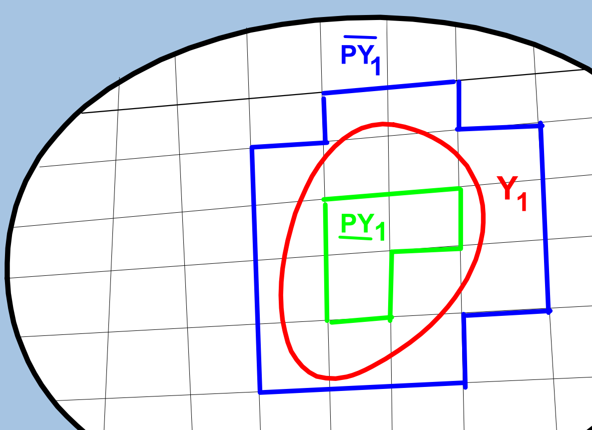
- $\underline{PY}_i$ is the union of all the elementary sets contained in Y_i

$$\underline{PY}_i = \{x / I_p(x) \subseteq Y_i\}$$

- $\overline{PY}_i$ is the union of all the elementary sets that contain any element of Y_i

$$\overline{PY}_i = \{x / I_p(x) \cap Y_i \neq \varnothing\}$$

Rough Set



Rough Set

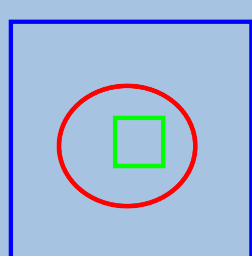
- We know that $\underline{PY}_i \subseteq \mathbf{Y_i} \subseteq \overline{PY}_i$
- The difference between the upper and lower approximation is the "**boundary** " of Y_i:

$$BN_P(Y_i) = \underline{PY}_i - \overline{PY}_i$$

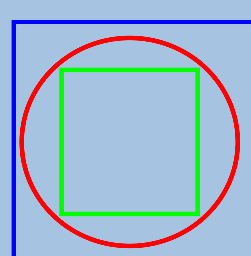
- $BN_P(Y_i)$ is the union of all the elementary sets with some elements belonging to Y_i and some belonging to the complementary of Y_i

Rough Set

- The size of $BN_P(Y_i)$ indicates the accuracy of approximation of set Y by $\underline{PY}_i$ and $\overline{PY}_i$
- The smallest $BN_P(Y_i)$ the better the approximation:



Bad approximation



Better approximation

Rough Set

- Every object in the set U is classified approximating Y_i by $\underline{PY_i}$:
  - If $x \in \underline{PY_i} \Rightarrow$ x is of class Y_i
- So the quality of approximation is $\gamma_p(Y)$
- We look for **subsets of attributes** R $\subseteq$ P, with the same quality of approximation that the whole set P:

$$\gamma_R(Y) = \gamma_P(Y)$$

Rough Set

- The subset with the smallest number of attributes are called **reducts**
- We choose the most interesting of them
- Its attributes are the most relevant ones (the other ones are discarded)

Rough Set

- A rule:

If ... Then ...

(If R1 >3 then the firm is healthy)

- Strength: number of objects covered by the rule

Linear Discriminant Analysis

- A classical multivariate technic concerned with separating distinct set of objects and allocating new objects to previously defined groups
- Many restrictive conditions: multivariate normality, equality of covariate matrix, ...
- If these conditions are violated the result may be questionable.

Data and Variables

- Sample: 72 spanish insurance firms; 36 failed and 36 non-failed (healthy)
- 3 models developed for 1, 2, and 3 years before the bankruptcy (Models 1, 2 and 3, respectivally)
- Training set: 75% of firms
- Test set: 25% of firms

Data and Variables

- 19 financial ratios for every firm (the attributes characterizing a firm)
- General financial ratios as well as specific ones (specific for evaluating insurance firms)

See 5 results

- Model 1 (one year before the bankrupcy):

```
  R13 > 0.68:
:...R9 <= 0.59: failed (14)
:   R9 > 0.59:
:   :...R17 <= 0.99: failed (3)
:       R17 > 0.99: healthy (3)
R13 <= 0.68:
:...R1 > 0.29: healthy (20/2)
    R1 <= 0.29:
    :...R2 > 0.04: failed (3)
        R2 <= 0.04:
        :...R6 > 0.64: healthy (3)
            R6 <= 0.64:
            :...R9 <= 0.85: failed (4)
                R9 > 0.85: healthy (4/1)
```

See 5 results

Model1

Evaluation on training data (54 cases):

```
     Decision Tree
    ----------------
    Size      Errors
      8    3(5.6%)   <<

    (a)   (b)    <-classified as
    ----  ----
     27          (a): class healthy
      3    24    (b): class failed
```

Evaluation on test data (18 cases):

```
     Decision Tree
    ----------------
    Size      Errors
      8    5(27.8%)   <<

    (a)   (b)    <-classified as
    ----  ----
      7     2    (a): class healthy
      3     6    (b): class failed
```

Rough Set results

- The values of the ratios are discretized and recoded to 1, 2 ,3 ,4.
- The quality of approximation with these ratios is 1
- We generate **reducts**: sets of attributes with the same quality of approximation as the whole set
- There are 229 reducts ⇒ the best of them is selected (for every model)

Rough Set results

- The reduct selected migth be the one with the most relevant attributes
- In our case the reduct have 5 attributes
- The rest of attributes are removed
- Decision rules are generated using the selected attributes

Rough Set results

- Model 1 ⇒ 27 rules
- Model 2 ⇒ 25 rules
- Model 3 ⇒ 25 rules

- Rules are used to classify the firms of the test set

Rough Set results

Model	Set of variables (reduct)	Number of decision rules	Correct classifications	
			"Healthy" firms	"Failed" firms
1	R3, R4, R9, R14, R17	27	77.78%	77.78%
			Total: 77.78%	
2	R1, R3, R4, R5, R17	25	75%	75%
			Total: 75%	
3	R2, R8, R11, R12, R18	25	57.14%	71.43%
			Total: 64.29%	

Linear Discriminat Analysis

- **A bad choice!!!**
  - Different covariate matrices
  - Not a multivariate normal distribution
  - Many outliers
  - Few degrees of freedhom
  - Poor results

Conclusions

- It seems that See5 performs slithgly better than Rough Set.
- Both methods performs much better than LDA.
- But Rough Set requieres a stronger intervention of the Decision Maker that must adjust some parameters

Conclusions

THE MORE DISCRIMINATORY RATIOS	DEFINITION
R1	Working capital/ Total Assets
R3	Investment Income/ Investments
R4	EBT*/ Total Liabilities
R9	(Capital +Reserves)/ Total Liabilities
R17	(Claims Incurred + Other Charges and Commissions)/ Earned Premiums

Conclusions

Model	Technique	Set of variables	Correct classifications	
			"Healthy" firms	"Failed" firms
1	See5	R13,R9,R17, R1,R2,R6	77.78%	66.77%
			TOTAL:72.22%	
	RS	R3,R4,R9, R14,R17	77.78%	77.78%
			TOTAL:77.78%	
	LDA	R1,R7	77.78%	44.44%
			TOTAL:61.11%	

Conclusions

Model	Technique	Set of variables	Correct classifications	
			"Healthy" firms	"Failed" firms
2	See5	R1,R13,R20, R7,R3	87.5%	75%
			TOTAL:81.25%	
	RS	R1,R3,R4, R5,R17	75%	75%
			TOTAL:75%	
	LDA	R12,R17	25%	75%
			TOTAL:50%	

Conclusions

Model	Technique	Set of variables	Correct classifications	
			"Healthy" firms	"Failed" firms
3	See5	R4,R19,R1	100%	57.14%
			TOTAL:78.57%	
	RS	R2,R8,R11, R12,R18	57.14%	71.43%
			TOTAL:64.29%	
	LDA	R4	57.14%	42.86%
			TOTAL:50%	