

Risk Analysis applied to Tax Evasion using data mining methodology⁽¹⁾

Analisi del Rischio Applicata alla Frode Fiscale Utilizzando Tecniche di Data Mining

Stefano Pisani

Agenzia delle Entrate – Ufficio Studi, Via del Giorgione 159, I – 00147 Roma,
Stefano.pisani@agenziaentrate.it

Paola De Sisti

So.Ge.I, Via Mario Carucci, 99 I- 00143 - Roma, pdesisti@sogei.it

Riassunto: Le tecniche di data mining sono state finalizzate all'individuazione delle frodi nella richiesta dei crediti d'imposta per "investimenti in aree svantaggiate". La scelta dello strumento è stata suggerita dai brillanti risultati ottenuti in ambiti analoghi (ad es. le frodi con carte di credito). Il modello associa a ciascun soggetto un punteggio che misura la probabilità che la richiesta sottintenda una frode. La sperimentazione condotta sul campo ha fornito risultati molto soddisfacenti (percentuale di successo dell'82%). Lo strumento si configura, quindi, molto indicato per perseguire altri tipi di frode, poste in essere tramite procedure automatizzate, che coinvolgono un'ampia platea di contribuenti (come le compensazioni di imposta o le richieste di rimborsi).

Keywords: Data Mining, Risk Analysis, Tax Fraud.

1. Introduction

The fight against tax evasion in Italy is particularly complex also due to the fact that the economic system is characterized by a larger number of enterprises in comparison to that of other industrialized Countries. The huge presence of small businesses characterizes Italian tax evasion as a mass phenomenon that, for this reason, is very expensive to fight. As a result there is a need for instruments able to block the evasion and deal with bulk loads of data. The proposed assignment is positioned in this line of research and applies the techniques of data mining to determine the fraud risk for a specific tax performance: collectible tax credits deriving from investments made in disadvantaged areas (hereafter tax credits).

2. Adopted procedures

A Sample of 54.517 subjects who, having already undergone a fiscal control on tax credits, are known to be tax evaders or not, were selected from a Population of 230.000 tax payers. The variable Target was then established, defining the evader as a subject who had in fact used the undue credit and, towards whom the Internal Revenue Agency could demand a restitution. In the exploratory phase several data banks of the Internal

⁽¹⁾ The work does not reflect the opinion of the member Institutions and solely commits the authors' responsibility.

Revenue Agency were integrated in order to obtain as much information as possible relative to the economic-fiscal behaviour of the tax payers included in the sample. From the initial data set a random sub-sample of 38.497 units (*training set*) was extracted and used to instruct the algorithm for the construction of the model to consent the determination of the tax evader profile risk, based on the target variable. Among the possible algorithms utilizable for the construction of the model it was decided to apply the one based on the *decision tree*, since it allows the visualization of a series of decisions easily transformable into operative indications for the control activities. The recursive partitioning algorithm used is CART and ID3 (Intelligent Miner-IBM). In the next phase the model was tested on that part of the sample excluded from the training phase (*test set* including 16.019 tax payers), evaluating the effectiveness of the adaptation by using the “Confusion matrix”. Since the object of the procedure was to use the risk analysis to perform ex-post controls, the model with the highest *precision* was chosen, to minimize the costs incurred from controls on the *false positives*. Finally the model was applied to the totality of the reference population (230.000 subjects). The leaf node renders, for the rule extracted, a prediction of the fraudulent class, that is the score attributed to subjects belonging to the class taken into consideration. This prediction is derived from the relationship between the number of subjects included in the fraudulent class divided by the total of subjects present in the leaf node. The Score assumes a value included between zero and one (maximum risk). Placing the score’s threshold at 0,85 in the test sample and considering the subjects above this value, it was demonstrated that the precision is equal to 0,82, a value considered acceptable; since on 100 controlled tax payers, the model tells us that 82 will be tax evaders and the error would be limited to 18 subjects.

3. Conclusion

The experiment provided extremely satisfying results, since a sector examination has produced a rate of 82% positive controls, equal to the probability estimated through the model. Furthermore, it has maximized the use of the information contained in the administrative data bases, obtaining indicators on the probability of being tax evaders solely from observing the information already available at the Internal Revenue Agency on each individual tax payer.

References

- Berger J. (1990) Robust Bayesian analysis: sensitivity to prior, *Journal Statistical Planning and Inference*, 25, 303-328.
- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*, Wadsworth.
- Cooper M. C., Milligan, G. W. (1988) The effect of measurement error on determining the number of clusters in cluster analysis, in: *Data, Expert Knowledge and Decision*, Gaul, W. & Shader, M. (Eds.), Springer, 319-328.
- Duda R. O. Hart, P. E. (1973) *Pattern Classification and Scene Analysis*, Wiley, New York