Clustering bankruptcy in Colombia: Comparative Analysis

Iván Andrés Trujillo
Juan Carlos Contreras

***Abstract*—Machine learning techniques are used broadly in international literature to predict bankruptcy, however in Colombian case there is not enough uses of the performance of these techniques. The main goal of this document is present a benchmarking about clustering techniques to predict bankruptcy in Colombia.**

A. *Introduction*

The financial distress is a term related with the situation of company insolvency, usually the term refer to the impossibility of payments, or its legal declaration of default, this situations have a negative impact in the economy, as reducing employment and raising prices. In summary, it could lead to reduce output and the improvements in living conditions of population.

I. Methodology

A. *Data and varaibles*

The data for this works is retrieved from the official web page of the entity in charge of oversight and control the companies in Colombia named in Spanish as Super Intendencia de Sociedades. The variables will be organized in a Python data frame to records the accounting statements, and the identification of the failure of each firm. The main variables used here, are the classical ratios that involve variables as current assets, current liabilities, operating revenue, net income, and other that are registered in the financial statements of the companies.

to identify the firms that were considered in bankruptcy we used *regular expressions* using keywords *liquid** and *reorganiza**.

We proposed the uses of two variables, to try clustering the bankruptcy in the Colombian case.

$$x_1 = \frac{\text{capital work}}{\text{total assest}} \tag{1}$$

$$x_3 = \frac{\text{Utility before tax}}{\text{total assest}} \tag{2}$$

B. *Outliers*

1) Shebyshev Theorem: Given the nature of data, we uses the shebyshev theorem to drop possible outlier values, that could bias our data.

$$1 - \frac{1}{k^2} \tag{3}$$

where k indicates the number of standard deviations, for this work we uses $k = 4$.

C. *Clustering techniques*

In the present work we uses the following clustering techniques:

- K-means
- K-means++
- Genetic K-means
- Fuzzy C-means
- DBscan

In summary, each methodology is a unsupervised technique and each one present drawbacks and limitations.

Although **k-means** is the cornerstone of these unsupervised techniques [**?**], however present some important limitations, for instance the crisp nature of its assignments, not capture no-spherical shapes. Due to its greedy nature, could be stuck in local optima and require a priori knowledge about the number of clusters. The another techniques could be tackle some of those problems; **Genetic K-means** tackle the greedy problem, reaching a global optima [**?**], **DBscan** find not spherical shapes and not require knowledge about the number of clusters [**?**], **Fuzzy c-means** allow to each pattern belong to all clusters with a degree of probability (tackle crisp nature) [**?**].

D. *Performance assessment*

Some authors point out an uneven importance of the type of errors in the assessment of performance, arguing that it is more profitable get higher sensitivity(True positive rate) than a major specificity (True negative rate), namely is better identify a higher numbers of firms that will failure than whose solvent firms classified correctly [**?**], [**?**].

We compare the performance of DBscan with the following clustering techniques; **k-means**, **k-means++**, **Genetic k-means**.

II. Result

According to the next table, the better model regarding the metric of performance is **k-means** followed by **Fuzzy C-means**.

	K-genetic	**K-means**	**K-means++**	**Fuzzy C-means**	**DBscan**
TPR	0%	88.3%	10.9%	63.4 %	32.5%

III. Discussion

According to the last table the models to consider are **K-means** and **Fuzzy- C-means**, now we are try to be a balance among precision and time complexity, according to big o notation the time complexity of **k-means** is $O(ncdi)$ and **fuzzy-c means** is $O(ndc^2i)$ where n represent the number of patterns, c the number of clusters, d number of features, i number of iterations [**?**], this could conclude that the better clustering technique either precision and complexity is k-means.

References

[1] J. MacQueen (1967). Some methods for classification and analysis of multivariate observations. Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1 (Univ. of Calif. Press, 1967), 281–297.

[2] Krishna, K., Murty, M. N. (1999). Genetic K-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 29(3), 433-439.

[3] Ester, M., Kriegel, H. P., Sander, J., Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

[4] Bezdek, J. C., Ehrlich, R., Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. Computers geosciences, 10(2-3), 191-203.

[5] Ghosh, S., Dubey, S. K. (2013). Comparative analysis of k-means and fuzzy c-means algorithms. International Journal of Advanced Computer Science and Applications, 4(4).

[6] Son, H., Hyun, C., Phan, D. Hwang, H. J. Data analytic approach for bankruptcy prediction. Expert Syst. Appl. 138, 112816 (2019).

[7] Alaka, H. A. et al. Systematic review of bankruptcy prediction models: Towards a framework for tool selection. Expert Systems with Applications vol. 94 164–184 (2018).