Clustering bankruptcy in Colombia: using K-means and Genetic K-means

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.**

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.

Genetic algorithm are attractive in some cases due some techniques require differential calculus, GA act a black box and only need evaluate the objective function. K-means no guaranteed a optimal solution, and GA could improve the solution, however, k-means suffer of a crisp nature given that in this algorithm is restricted each point to belonging to a only one centroid, thus could produce empty centroids, this problem could be tackled with C-fuzzy means.

I. Background

Clustering is a technique used to understand the structure of data behind bankruptcy, the following works are used clustering technique to predict and agglomerate bankruptcy [2], [3], [5], and some works document the outperform of GKA over K-means [4].

II. Data

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

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

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

A. 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 $k = 4$.

III. Methodologies

IV. K-means

A. K-means++

Given the dependency of the result to the initial centroids, this algorithm tackle this problem.

1) Numer of clusters: We define a number of two cluster a priori.

B. K-means

In a sample of n patters and d features we can define $k = 1, 2, ..., K$ as the number of clusters, the centroid c_k will be defined as $c_k = (c_{k1}, c_{k2}, c_{k3}, ..., c_{kd})$, note that each component correspond to each feature, specifically:

$$c_{kj} = \frac{\sum_{i=1}^{n} w_{ik} x_{ij}}{\sum_{i=1}^{n} w_{ik}} \tag{4}$$

in the last equation $w_{ik} = 1$ if $i - th$ sample belong to the cluster k and $w_{ik} = 0$ in otherwise.

in the following equation the matrix $W = [w_{ij}]$ is composed of value or belonging to each cluster $w_{ij} =\in [0, 1]$.

the within cluster variation of the k cluster is:

$$S^k(W) = \sum_{i=1}^{n} w_{ik} \sum_{j=1}^{d} (x_{ij} - c_{kj})^2 \tag{5}$$

therefore the total within cluster variation is:

$$S(W) = \sum_{k=1}^{K} S^k \tag{6}$$

the goal is find W^* that minimize $S(W)$.

Although is a practical algorithm k-means is sensitive to the initial centroids and not guaranteed reach a global optima [1].

V. Genetic K-means

Given that k-means stuck in the first stable sets of centroids (given the greedy nature of the algorithm) we uses a genetic approach to tackle this problem. GKA Proposed by [1] and inspired in natural selection process of the genetic evolution, in summary the algorithm consist in initialize a random population and evolve to generations.

The search space is all W in which $\sum_{i=1}^{K} w_{ij} = 1$. the solution could be represented as:

$$x = (c_1, c_2, ..., c_k) \tag{7}$$

Note here that c_k is defined in (4).

A. Genetic k means details

1) Chromosome representation: given a number fixed of cluster in this case $k = 2$.

Fig. 1. Chromosome representation

The last figure indicate that the first individual belong to cluster A , the second to B and the last n belong to A.

2) Random population: Generate T samples of chromosomes of the same longitude n.

B. Natural selection

1) Fitness function: The goal is evaluate the fitness function $f(x)$ in all possible candidates therefore,

$$f(x) = \frac{1}{S(W)} \tag{8}$$

2) Biased selection OR selection operator: Select possible chromosomes as parents.

$$P_{x_i} = \frac{f_i(t)}{\sum_j^n f_j(t)} \tag{9}$$

after of get we can choose a random number among 0 and the sum of total fitness in population, this strategy is denominated wheel strategy.

3) Chromosomal crossover: the offspring born choosing a random locus (position in chromosome), in pairs of parents, and after concatenate for instance α_a and α_b and the locus is equal to π then the offspring will be $\alpha_a[:\pi]$ concatenate with $\alpha_b[\pi:]$.

C. Mutation

Adding random mutations produce a search of solutions, this process could be summarized as change the allele in a random locus.

D. Stopping criteria

the best possible fitness score not is known, then a early stopping criteria could be uses a elbow method when there are not substantial improvements in the fitness in a given generation.

VI. Assessments of quality

Given that we know the true label we compare the distribution of each predicted member cluster regarding with the true distribution, to asses the quality of the overall classification we calculate the χ^2 test.

VII. Results

According to the table we can see that the GKA proposed a more optimal and balanced solution than K-means.

			Bankruptcy		
		Overall	**0**	**1**	**P-Value**
n		762	497	265	
GKA, n (%)	**0**	268 (35.2)	202 (40.6)	66 (24.9)	<0.001
	1	494 (64.8)	295 (59.4)	199 (75.1)	
K-means, n (%)	**0**	50 (6.6)	19 (3.8)	31 (11.7)	<0.001
	1	712 (93.4)	478 (96.2)	234 (88.3)	

However according to the results of p values both models present different expected and real distribution of categories. This means that even thought GKA proposed more optimal solution not is enough uses the variables x_1 and x_2 as unique possible features.

References

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

[2] Martin, A., Gayathri, V., Saranya, G., Gayathri, P., Venkatesan, P. (2011). A hybrid model for bankruptcy prediction using genetic algorithm, fuzzy c-means and mars. arXiv preprint arXiv:1103.2110.

[3] Prusak, B. (2018). Review of research into enterprise bankruptcy prediction in selected central and eastern European countries. International Journal of Financial Studies, 6(3), 60.

[4] Pizzuti, C., Procopio, N. (2016, October). A K-means based genetic algorithm for data clustering. In International Joint Conference SOCO'16-CISIS'16-ICEUTE'16 (pp. 211-222). Springer, Cham.

[5] Le, T., Le Son, H., Vo, M. T., Lee, M. Y., Baik, S. W. (2018). A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset. Symmetry, 10(7), 250.