ATLANTIS PRESS

Atlantis Studies in Uncertainty Modelling, volume 1

11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)

Pattern Classification on Complex System Using Modified Gustafson-Kessel Algorithm

Miguel Khatounian Filho[a], Leo Koki[a] and Renato Aguiar[a]

[a]Department of Electrical Engineering, Centro Universitário FEI, S.B. CAMPO, Brazil
miguelkf@outlook.com, leo.lks@hotmail.com.br, preraguiar@fei.edu.br

Abstract

This work has been focused in the application of Gustafson- Kessel Algorithm in a complex system through a methodology proposed. The complex system here considered will be the financial market. So, the main objective of this paper is to classify objects in two patterns: winner and loser. The methodology is based on application of a method of clustering called Modified Gustafson-Kessel (MGK) in some open companies of the transportation sector and energy sector. Results shows that the use of MGK can better separate the promising actions from the non-promising ones with more precision due to its covariance matrix that can be change for generate the best separability among clusters. This produces a new tool for analysis of the dynamic of stock market with the main aim of given support to investor in make decision.

Keywords: Fuzzy c- Means Algorithm, Gustafson- Kessel Algorithm, Stock Classification, Pattern Recognition

1 Introduction

A complex system can be understood as a collection of many interdependent parts that contain an interation with each other through nonlinear collaboration. As example of complex system can be mentioned the human brain, the weather, economy and financial markets. In fact, the financial market is a non-linear and time-variant system, and can belongs to regions of stability and instability influenced by many variables such that: optimism and pessimism of the investor, news, financial indicators, etc. Thus, the financial market is a complex system and requires many mathematical models to understand it and to help investors in the risk analysis, financial forecast, make decision, formation of portfolio. There are several papers that deal with these issues, developing models for portfolio formation as can be seen in [1], [2] and [3], among others.

Obviously, for this complex system, several tools were developed in theories of administration and economics with the aim of modeling this market [4], [5]. However, a few years ago, the techniques used in engineering, especially those related to artificial intelligence, have been widely applying in the financial market. In [3], the authors developed stock classification model using the fuzzy clustering algorithm called fuzzy c-means. Among other studies, it is worth mentioning [6] where the theory of genetic algorithm was utilized in the financial market, also in [7] the authors use neural networks.

In this way, in this work will be utilized, for the stock classification of the financial market, Gustafson-Kessel Modified Algorithm (MGK). Since the FCM was already used for this purpose, as can be seen in [3], the main objective here is to apply the same technique proposed in [3], using GKM algorithm. The methodology adopted will be the same as in [3], using real data of stock market from 2011 to 2016.

It is expected that application of GKM algorithm in stock market can produce better results than FCM algorithm, since GKM seems to have more flexibility in regarding to FCM, which can significantly improve the selectivity of stocks and the accuracy of the results.

2 An Overview

Clustering analysis is the formal study of algorithms and methods for grouping or classification of objects. The fuzzy c-means [8] and Gustafson-Kessel Modified [9] algorithms are very useful tools to accomplish this grouping, defined by the similarity between some features related to these objects.

There are basically two grouping techniques: hierarchical and partitional. The hierarchical is commonly used in the biological sciences and represented by a tree based in features of the objects, known as a dendogram, which allows visualization of the groups and their relationships. On the other hand, partitioned grouping is based on the representation of the data through its characteristics arranged in scattered form. To facilitate this study, a pattern matrix containing the

database is formed, where each of the ***n*** objects is represented by a set ***d*** of features. Each row of the matrix defines an object and each column is associated with a specific feature, thus forming an array of dimensions n x d [10], [11].

The goal is to determine a number of clusters, so that objects within a group are more similar than objects in other groups. This similarity is based on the distance between objects with respect to a center (called the cluster center), that is, as smaller the distance between an object and a center, higher will be the membership degree of this object in regarding to this center.

As an example, two hypothetical groups obtained through of this grouping is presented at Figure 1.

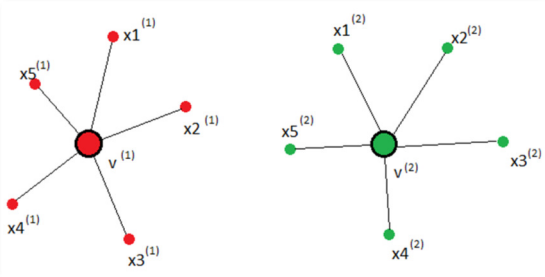

Figure 1 – Hypothetical Clusters of 2 dimensions

As can seen in Figure 1, each group or cluster is represented by a center V, and surrounded by some objects. The center V can be understood as a reference to other objects. That is, the center is responsible for representing all the objects present in the cluster.

It is worth remembering that in this traditional grouping, each object belongs exclusively to a single group. In the case of a fuzzy grouping, each element belongs to more than one group simultaneously but with different degrees of similarity. The distances of each object in regarding to each center will determine these degrees of similarity, so that the smaller the distance of an element in regarding to a center, greater the degree of similarity between this object and the corresponding group.

The clustering algorithms based on fuzzy logic theory are direct modifications of the partitioning algorithms with the quadratic error criterion.

The fuzzy set theory [12] has as one of its main characteristics the fact of allowing intermediate values to be defined between the conventional values as true or false, hot or cold, etc.

According to [13], fuzzy logic is a tool able of capturing vague information, generally described in natural language, and converts them into a numerical, easy-to-manipulate format.

Thus, analyzing the variables from this perspective, unlike the classical set theory, where objects belong or not to a set, that elements have similarity degrees associated with each one of the groups.

We can describe the membership grade as a measure of similarity, proximity or compatibility between two elements. Mathematically its definition is formulated as follows:

Let X be a collection of objects $X = \{x_1, x_2, \dots, x_n\}$, $A = \{A_1, A_2, \dots A_p\}$ a subgroup of X. and $0 \leq \mu_i(x_j) \leq 1$, $i = 1,2, \dots p$ e $j = 1,2, \dots n$, such that all $j = 1,2, \dots n$, the equality $\sum_{i=1}^{m} \mu_i(x_j) = 1$. So, $\mu_i(x_j)$ is called the membership grade (or degree of similarity) of the element x_j in regarding to the subset C_i.

Among the techniques for grouping or fuzzy classification of elements into subsets of a given set, fuzzy c-means (FCM) algorithm is very efficient, as well as being used as a framework for the formulation of other algorithms.

The algorithm that implements FCM forms its partitions by minimizing the following objective function:

$$min \sum_{i=1}^{p} \sum_{j=1}^{n} \left[\left(\mu_i(x_j) \right)^m \left\| x_j - c_i \right\|^2 \right] \qquad (1)$$

where x_j representes the elemento j, c_i representes the center of clusters and $m \in [1, \infty)$ is the weighting exponent.

It's an optimization problem and can be solved through of equations (2) and (3) and following steps 1 to 4 [11].

$$c_i = \frac{\sum_{j=1}^{n} \left(\mu_i(x_j) \right)^m x_j}{\sum_{j=1}^{n} \left(\mu_i(x_j) \right)^m} \quad i = 1,..p \qquad (2)$$

$$\mu_i(x_j) = \frac{\left(\frac{1}{\left\| x_j - c_i \right\|^2} \right)^{\frac{1}{m-1}}}{\sum_{k=1}^{n} \left(\frac{1}{\left\| x_j - c_i \right\|^2} \right)^{\frac{1}{m-1}}} \qquad (3)$$

$\boldsymbol{i = 1,..p \quad j = 1,2, \dots n}$

Step 1: Start a membership grades matrix, so that $\sum_{i=1}^{p} \mu_i(x_j) = 1$

Step 2: Calculate the centers through equation (2);

Step 3: Recalculate, through equation (3), the new membership grades matrix using the center vectors obtained in step 2;

Step 4: Repeat steps 2 and 3 until the value of the objective function, equation (1), no longer decreases, according to the precision adopted.

However, in 1979 a work was published proposing a modification for the traditional Fuzzy c-Means Algorithm (FCM). The modification described in this work was entitled Gustafson-Kessel (GK), due to name of its authors [15].

In FCM algorithm is utilized Euclidean distance between each object and each center, while that in GK

algorithm is utilized the Mahalanobis distance, which implements a covariance matrix among the attributes available in the database. This matrix has the function of calculating the relation between the different properties, in order to allow better flexibility in the formation of each cluster [14].

The covariance matrix allows the Gustafson-Kessel algorithm to find clusters of independent geometric shapes, that is, each group has its own dimensional features. Therefore, the results generated by the GK algorithm can be better than FCM algorithm.

However, some numerical problems occur frequently when GK is applied in its standard form and the number of samples is small or when data within a cluster is practically linearly correlated.

Considering this limitation of the GK algorithm, a method was proposed to overcome this problem, which was called Modified Gustafson-Kessel (MGK). Such modification can significantly improve the performance of the GK algorithm [9]

The GKM algorithm also follows the principle of optimization of a similar function to that proposed by the FCM algorithm, given by:

$$J = \sum_{i=1}^{K}\sum_{k=1}^{N}(\mu_{ik})^m \, D^2{}_{ikAi} \qquad (4)$$

$$\text{with } D^2_{ikAi} = \left(z_k - v_i^{(l)}\right)^T A_i \left(z_k - v_i^{(l)}\right)$$

being the distance between the element I and cluster k, weighted by matrix A_i

Where N represents number of companies, K represents the number of clusters, μ_{ik} is the membership grade of each object (company) k in regarding to each cluster and D is a distance, which produces clusters with different shapes. The equations (5), (6) and (7) are the base of the algorithm [9].

$$v_i^{(l)} = \frac{\sum_{k=1}^{N}(\mu_{ik}^{(l-1)})^m \, z_k}{\sum_{k=1}^{N}(\mu_{ik}^{(l-1)})^m} \quad i = 1,2,..K \qquad (5)$$

$$F_i = \frac{\sum_{k=1}^{N}(\mu_{ik}^{(l-1)})^m (z_k - v_i^{(l)})(z_k - v_i^{(l)})^T}{\sum_{k=1}^{N}\left(\mu_{ik}^{(l-1)}\right)^m} \qquad (6)$$
$$i = 1,2,...K$$

$$F_i^{new} = (1-\gamma)F_i + \gamma det(F_o)^{1/n} I \qquad (7)$$

Where $\gamma \in [0,1]$ is a control parameter and F_o is the covariance matrix of the entire data set. Depending on the value of γ, the clusters are forced to become more or less similar in their form.

Then eigenvalues λ_{ij} and eigenvectors ϕ_{ij} are obtained from F_i^{new}. Find $\lambda_{i\,max} = max_j \, \lambda_{ij}$ and set $\lambda_{ij} = \frac{\lambda_{i\,max}}{\beta} \quad \forall j, resulting \; \frac{\lambda_{i\,max}}{\lambda_{ij}} > \beta$, and obtain the reconstruct F_i by

$$F_{i=} = [\Phi_{i1} \ldots \Phi_{in}] dig(\lambda_{i1}, \ldots, \lambda_{in}) [\Phi_{i1} \ldots \Phi_{in}]^{-1} \qquad (8)$$

Briefly the GKM algorithm can be executed thought the following steps:

Step 1: Initialize a merbership grades matrix, so that $\sum_{i=1}^{K} \mu_{ik} = 1$.

Step 2: Calculate the centers of the each cluster through equation (5).

Step 3: Calculate the cluster covariance matrices through equations (6) - (8).

Step 4: Calculate the distances through equation (9) given by.

$$D^2_{ikAi} = \left(z_k - v_i^{(l)}\right)^T \left[\rho_i det(F_i)^{\frac{1}{n}} F_i^{-1}\right] \left(z_k - v_i^{(l)}\right) \qquad (9)$$
$$i = 1,2,...K \qquad k = 1,2,...N$$

Step 5: Update the membership matrix through equation (10) using the distances obtained in step 4.

$$\mu_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{K}\left(\frac{D_{ikAi}}{D_{jkAj}}\right)^{\frac{2}{m-1}}} \qquad (10)$$
$$i = 1,2,...K \qquad k = 1,2,...N$$

Step 6: Repeat steps 2- 5 until objective function shown in equation (4) not more decrease.

The application here considered has as fundamental base the modified Gustafson- Kessel algorithm and possess two main stages: pattern recognition and stock classification.

3 METHODOLOGY

In this section will be introduced the methodology used for stocks classification using Modified Gustafson- Kessel Algorithm (MGK). So, the methodology is based on techniques of pattern recognition and has as fundamental base the GKM algorithm, where the features or information of each stock are financial indexes of some companies. These indexes were collected quarterly from the Economatica database [16] from the first quarter of 2011 to the third quarter of 2016.

Four indexes were selected, and considered the most significant because they have a strong relationship with the financial return of the stocks: Net Margin,

Price on Profit, Price on Net Asset Value and Net Debt on Shareholders' Equity [3].

In this first stage, the aim is to separate the data into two groups: a promising group, which will produce a greater financial return, and an unpromising group, which will produce a lower financial return. In this first stage, was considered the total period from 2011 to 2014.

During this stage of pattern recognition, the total period was divided into quarterly subperiods, t = 1,2, ..., i. For the quarter t, GKM algorithms was applied, in order to separate the stocks into two different groups.

In each quarter t, the algorithms was applied to the pattern matrix M, of dimension n x d, where each *n* rows corresponds to one company, and each *d* columns corresponds to the financial indexes related to the company. Then, after some iterations, two groups were formed and, for each cluster was calculated the average financial return produced. The group that produced a higher average financial return is called here a good group, classified in the promising stock group, and the one that produced the lowest average financial return is called bad group and classified in the group of non-promising stocks. Figure 2 summarizes this first stage for a quarter *t*.

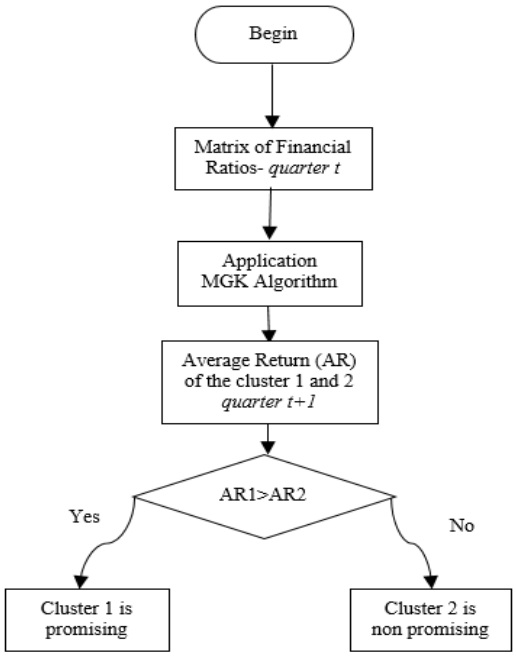

Figure 2 – Stocks Pattern Recognition

This procedure was applied on 15 quarters, from 1^{st} quarter of 2011 to 3^{rd} quarter of 2014, generating two patterns: a promising pattern and another pattern called no promising.

In the second stage, with the patterns founded in the first stage, the stocks were classify from 1^{st} quarter of 2015 to 3^{rd} quarter of 2016. Each quarter considered in the first stage produced two center vectors: a promising center vector e an unpromising center vector. In other words, there are 15 winner center vectors (v_w) and 15 loser center vectors (v_L) generating a matrix of center vectors *V* given by:

$$V = \begin{bmatrix} v_w^1 & v_L^1 \\ \vdots & \vdots \\ v_w^{15} & v_L^{15} \end{bmatrix}$$

In this matrix *V* was applied the GKM algorithm and obtained only one promising center vector and one unpromising center vector, which are called winner center vector and loser center vector, respectively.

After this, considering the period from 2015 to 2016, in each quarter t was calculated Mahanalobis distance between each stock and each center of centers through equation 9 and consequently the membership grade correspondent through equation 10. The stocks with higher membership grades in regarding to winner center vector were classify as winner stocks and stocks with higher membership grades in regarding to loser center vector were called as loser stocks. This suggests two portfolios: a winner portfolio and a loser portfolio.

The figure 3 shows the procedure adopted in the second stage.

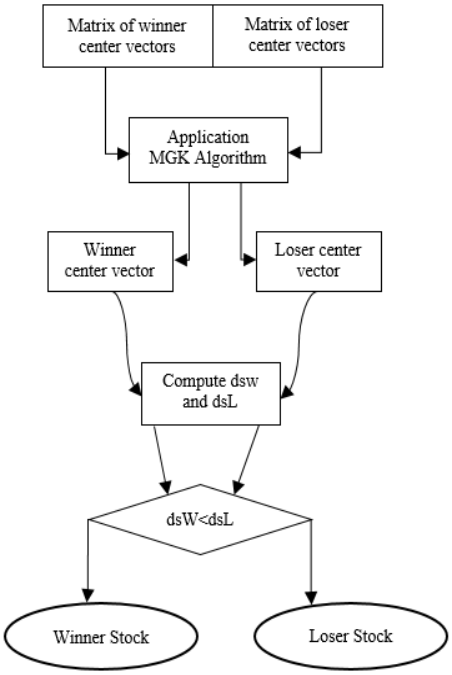

Figure 3: Stocks Pattern Classification

Where dsW is the distance between a stock and winner center vector and dsL is the distance between a stock and winner center vector.

This procedure was applied in each quarter from 2015 to 2016 producing in each quarter a winner portfolio and a loser portfolio.

4 RESULTS

For the application of the methodology presented above the two sectors of stocks were considered separately: stocks of the transportation sector and energy

sector. This make decision to analyze the sets separately because the joint analysis of stocks of different sectors is not recommended when financial indices are used (Matarazzo, 2010).

In order to obtain the results, Table 1 shows the application of the GKM algorithm in data of the stocks of the energy sector in the 1^{st} quarter of 2011, while table 2 shows the financial return of each stock in the 2nd quarter of 2011.

Companies		**Membership Degrees**	
		μ_1	μ_2
AES Elpa	1	0,973	0,027
Celesc	2	0,920	0,080
Cemar	3	0,010	0,990
Cemig	4	0,002	0,998
Cesp	5	0,640	0,360
Coelba	6	0,009	0,991
Coelce	7	0,498	0,502
Copel	8	0,917	0,083
CPFL Energia	9	0,086	0,914
Eletrobras	10	0,983	0,017
Eletropaulo	11	0,970	0,030
Energias BR	12	0,931	0,069
Energisa	13	0,116	0,884
Engie Brasil	14	0,244	0,756
Equatorial	15	0,010	0,990
Light S/A	16	0,197	0,803
Taesa	17	0,657	0,343
Tran Paulist	18	0,797	0,203

Table 1 – Membership Matrix 1^{st} quarter 2011

Companies		**Financial Return [%]**
AES Elpa	1	12,32
Celesc	2	-4,86
Cemar	3	8,37
Cemig	4	7,37
Cesp	5	0,38
Coelba	6	0,00
Coelce	7	9,38
Copel	8	-4,97
CPFL Energia	9	-1,76
Eletrobras	10	-1,71
Eletropaulo	11	8,63
Energias BR	12	-1,46
Energisa	13	27,66
Engie Brasil	14	0,52
Equatorial	15	8,16
Light S/A	16	10,86
Taesa	17	28,52
Tran Paulist	18	-0,34

Table 2 – Financial Return 2^{nd} quarter 2011

Table 1 shows nine companies with a higher membership degree in regarding to cluster 1 and nine companies with a higher membership degree in regarding to cluster 2 on 1st Quarter 2011. Table 2 shows the financial return of each company in the second quarter of 2011, so that the financial return of each cluster was calculated: for cluster 1 2.945% and for cluster 2 7.803%. Then, cluster 2 produced an average financial return greater than cluster 1 and can be called a good or promising cluster or winner cluster.

This procedure was repeated for all quarters, from the 1st Quarter of 2011 to the 3rd Quarter of 2014, obtaining, in this way, the promising group and the no promising group for each quarter.

Figure 4 shows the financial return generated by the promising and unpromising groups during these fifteen quarters analyzed using GKM algorithm

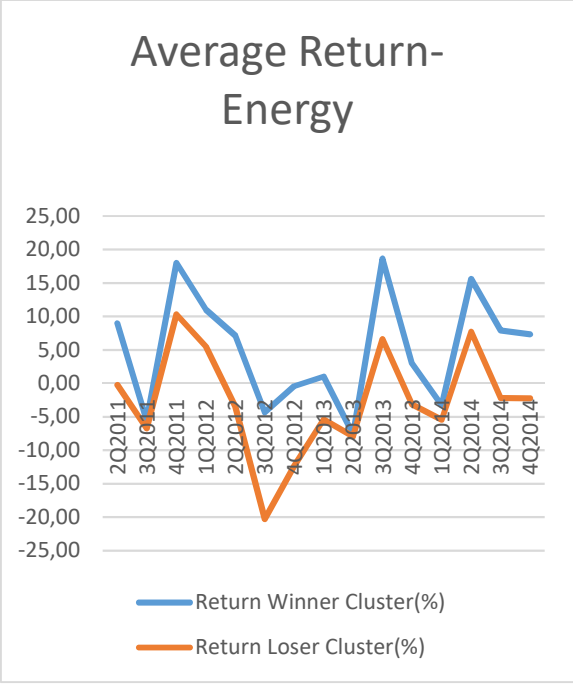

Figure 4 – Average Returns – Energy

Figure 4 reveals that utilizing Modified Gustafson-Kessel algorithm is utilized, it's possible obtain promising clusters with average financial return considerable great than unpromising cluster. This indicate that

Modified Gustafson- Kessel algorithm allows separate with more precision the promising stocks.

As can be seen, the separability power between promising and unpromising stocks through the application of GKM algorithm is undeniable.

So, the pattern generated by the GKM algorithm shows to be more precise and more selective because of its flexibility. These data are used in second stage.

A) Second Stage: The application of GKM algorithm in each one of these 15 quarters produced a promising center vector (winner)- vw and a unpromising center vector (loser)- vL. The result is a center matrix V given by:

$$V = \begin{bmatrix} v_w^1 & v_L^1 \\ \vdots & \vdots \\ v_w^{15} & v_L^{15} \end{bmatrix}$$

Here, the aim is to obtain only a promising center vector called winner center and an unpromising center vector denominated loser center. In this sense, the winner center was obtained as follow: in the matrix of center vector V was applied the GKM algorithm in the set winner center vectors, producing two cluster and each one cluster possess a center. These two centers is called here winner center of centers. After this, tests were done for identify the best winner and the same procedure was done to set loser center. The table xx shows the winner center which represents all winner center vectors and loser center which represents all loser center vectors. The table 3 shows the winner and loser centers to energy sector

	DL/PL	P/L	P/VPA	ML
Winner Center	76.50	14.82	1.60	11.63
Loser Center	62.05	-2.37	1.42	10.71

Table 3: Winner and loser centers- Energy

The classification will be done using the Mahalanobis distance. The Mahalanobis distance between each stock and each of the winner and loser centers is computed. This distance, contained in GKM algorithm, has a relationship with membership degrees, so that if the distance between a stock and winner center is smaller than distance between this stock and loser center, then such stock is more similar with winner center and can be classified as winner stock. This procedure was done for some stocks of energy sector from first quarter of 2015 to 3º quarter of 2016. So, after classification two portfolios are formed: winner portfolio and loser portfolio. The stocks closer winner center, what means stocks with greater membership degree in regarding to winner center, belong to promising portfolio.

The figures 5 shows the average return of the winner portfolio together with average return of loser portfolio for energy sector.

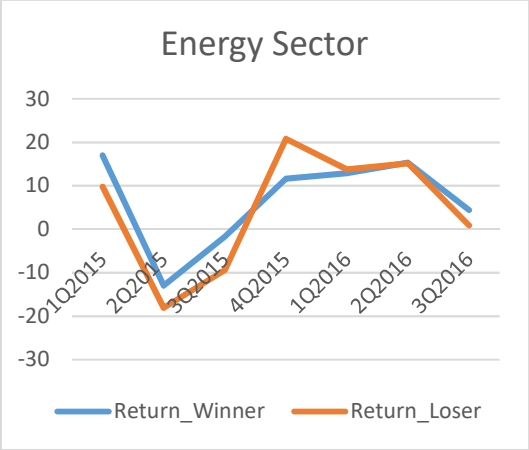

Figure 5 – Average Returns –Classification Energy

Considering this total period, the average return of winner portfolio is 46.67% against 33.06% of loser portfolio.

The same methodology was applied to companies of transportation sector. The results followed the same trend obtained for the energy sector and can be seen in figure 6. Table 4 shows the winner and loser centers to transportation sector.

	DL/PL	P/L	P/VPA	ML
Winner Center	124.12	19.94	3.42	5.56
Loser Center	148.79	27.19	3.62	8.45

Table 4: Winner and loser centers – Transportation

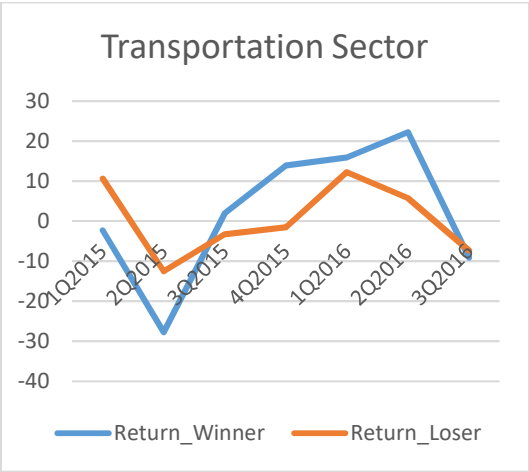

Figure 6 – Average Returns – Transportation

Considering this total period, the average return of winner portfolio is 15.05% against 4.08% of loser portfolio.

This classifier is more precise, more flexible and with more stock selectivity, with great potential for medium-term financial returns when compared to the one

presented in [3]. This proposal, together with other analyzes, may assist an investor in making investment decisions at any time.

5 CONCLUSION

This study investigated the efficiency of the identification of patterns of open companies using the GKM algorithm. It was found that, with GKM algorithm - usually used in image recognition, speech recognition, among other things – it's possible to recognize promising and non-promising patterns for stocks in the financial market. Although the FCM was used in previous work, the efficiency of GKM in the stock market has been investigating in this work. Differently of the FCM, the GKM is more precise due to its flexibility in regard to distance considered. The FCM, due to Euclidean distance, can lost some companies promising.

The results obtained here for companies of the energy and transportation sectors have shown that: in the most of quarters, GKM algorithm is designed to separate the promising stocks from the non-promising ones with more efficiency, finding a better optimal local point. In addition, the total financial return for the winner portfolio is very significant.

The algorithm GKM gets to separate the promising stocks from non- promising stocks forming a portfolio with maximized earn. In this way, can be concluded that the use of GKM is efficient, reliable and more flexible in application to pattern recognition of stocks of stock market, with the ability to generate more promising short-term results and in medium term. Consequently, this application has the potential to generate greater gains for the investor. The model here presented has as main aim to given support to investor in make decision in a complex, unstable environment called financial market.

References

[1] Dourra H.; Siy P. (2002). Investment using technical analysis and fuzzy logic. Fuzzy sets and systems, p. 221-240, April 2002.

[2] Kuo, R. J.; Chen, C.H.; Hwang, Y.C. (2001). An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artificial neural network. Fuzzy sets and systems, v. 118, p. 21-45.

[3] Aguiar, R.A.; Sales, R.M. Stocks Classification Using Fuzzy Clustering. In: The 2004 International Conference on Artificial Intelligence, 2004, Las Vegas. Post-Conference Proceedings/books. Las Vegas: CSREA Press, 2004. v. 1. p. 249-255.

[4] Lima, M. L. Um Modelo de Predição de Bolsa de Valores Baseado em Mineração de Opinião, Dissertação de Mestrado – Universidade Federal do Maranhão, 2016.

[5] Hayashi, A. H. Processo para Predição de Preços das Ações no Mercado Financeiro com Uso de Big Data – Instituto de pesquisa Tecnológica do Estado de São Paulo,2017.

[6] Kuo, R. J.; Chen, C.H.; Hwang, Y.C. (2001). An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artificial neural network. Fuzzy sets and systems, v. 118, p. 21-45.

[7] Wong, F.S.; Wang, P.Z. (1991). A stock selection strategy using fuzzy neural networks. Neurocomputing, v. 2, p. 233-242.

[8] Dunn, J.C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybernetics, V. 3, N. 3, pp. 32-57.

[9] Babuska, R.; Van der Veen, P.J.; Kaymak, U.. Improved Covariance Estimation for Gustafson- Kessel Clustering. Proc. of the IEEE International Conference on Fuzzy Systems, Honolulu, USA, 2002.

[10] Zimmermann, HJ. Fuzzy Set Theory—and Its Applications. 4. Ed. Springer Science & Business Media, 2011.

[11] Bezdek, J. and Pal, S.K. (1992). Fuzzy Models for Pattern Recognition. IEEE, pp. 88-94.

[12] Zadeh, L. A. Fuzzy Sets, Information and Control, v. 8, 1965.

[13] Wagner, A. Extração de Conhecimento a partir de Redes Neurais aplicada ao problema da Cinemática Inversa na Robótica. Dissertação de Mestrado. Universidade do Vale do Rio dos Sinos, 2003.

[14] Bezdek, J. C. et al. Fuzzy models and algorithms for pattern recognition and image processing. Springer, 2005.

[15] Gustafson D.E. and Kessel W.C.. Fuzvy clustering with a fuzzy covariance matrix. In ***Proc.IEEECDC,*** pages761-766, San Diego, CA, USA, 1979.

[16] Economática Ltda (2002). Software of support, www.economatica.com.br, accessed on 2002