A New Look on Classification of Spanish Mutual Funds[1]

Working Paper in progress: First Draft

Please, do not quote without permission of authors

David Moreno [a, c], Ignacio Olmeda [a, b], and Paulina Marco [a, d]

[a] *Laboratorio de Finanzas Computacionales*
[b] *Dpto. de Ciencias de la Computación*
[c] *Dpto. de Fundamentos de Economía e Historia Económica*
Universidad de Alcalá
[d] *Dpto. de Finanzas Empresariales*
Universidad de Valencia

Abstract: In this paper, we apply a nonlinear methodology (Self-Organizing Maps and k-means algorithm) to classify mutual funds. This methodology allows us to identify which mutual funds did not follow the investment objectives established from its official category. We also propose an alternative classification, which has a lower rate of misclassified mutual fund. It achieves better performances in terms of return/risk, with a lower number of assets, which permits increase the final utility of particular investors. We find that in average 60% of Spanish mutual funds are misclassified, according with their financial attributes. It must be taken into account that grouping mutual funds correctly may have several implications from ex-post and ex-ante benchmark and performance measures.

Key words: Self-Organizing Map (SOM), Artificial Neural Network (ANN), Performance Measures, Clustering Methods.

1. Introduction

Mutual funds have become the mainly investment instrument for particular investors. In the last ten years, they registered a huge growth, both qualitative and quantitative. Accordingly to this fact, in June, 1999, an exhaustive and detailed classification was established by the CNMV ("*Security Market National Commission*" in Spain) and INVERCO (*"Spanish Association of Collective Investments Institutions"*).

[1] Corresponding author: Researcher David Moreno, Dpto. Fundamento de Economía e Historia, Universidad de Alcalá, Alcalá de Henares (Madrid), SPAIN. Email: **jesusdav@eresmas.com**

We must remark that a correctly settled classification could be beneficial for making investments decisions. But, what is a good classification?. We think that this one that:

1. Reflects the mutual funds financial attributes (return-risk)
    - Helping particular investors to diversify his welfare, improving in return-risk terms, and
    - Avoiding the tendency to "gaming" of benchmark from mutual fund managers.

In this paper, we try to shed some light about the above questions, in order to improve, in the possible measure, the actual official classification and their consequences for investors.

Thus, in the present study, we apply a nonlinear methodology to classify mutual funds, which picks up all nonlinear patterns and relations between the whole of funds attributes. This method allows us to identify the mutual funds misclassified, permeating us to reclassify them. Besides, it achieves better diversification performances with a lower number of assets, which permits increase the final utility of particular investors.

The results of this paper have important implications, both for academics studies and for partitioning. For example, knowing the correct group in which a mutual fund should be classified, has an extremely relevance for performance persistence studies (e.g. Brown and Goetzmann, 1995), that's means that those mutual fund which were winner during the last time period, continuing been winners in the next one. For knowing it, is indispensable to know the group in which is really allocated each fund according with its financial characteristics, instead of; from the declaration that fund managers did when the mutual fund was born. Moreover, grouping mutual funds correctly may have several implications from ex-post and ex-ante benchmark and performance measures construction and evaluation (see Grinblatt and Titman, 1994).

From a particular investor point of view, a mutual funds misclassification has also a huge relevance, given that they must allocate their savings into those funds that better satisfy their financial necessities. Then, it is extremely relevant to know which characteristics, in terms of risk-return, have truly each mutual fund. As to get the objective of diversification, the investor select between mutual funds of different categories, it is strictly important that all they are correctly classified. If, according as our results, it does not happen, the investor is going to pick up the erroneous mutual funds. This fact could cause a suboptimal asset allocation and, as a result, a reduction for the generality of the Spanish economy.

The mainly characteristics that the ideal classification should have is objectivity and consistency. As Brown and Goetzmann (1997) say:

" *(...) Because management styles are so widely used as the basis from performance measurement and compensation, there is a great need for stylistic classifications that are objectively and empirically determined, consistent across managers and related to the manager's strategy. The objectivity is important because of the moral hazard inherent in allowing managers to self-report their styles, without objective verification. The consistency is needed for purposes of performance comparison*".

Objectivity must be require to avoid the manager's mutual funds could have the temptation of declaring an objective investment and realizing some other. They would do it following to climb up their position into the public rankings. If it is not avoid, managers' funds could take a more risky positions that they declared initially, following to increase their performance evaluations. As Marathe and Shawky (1999) comment, this behaviour can be especially important from the smaller company's mutual funds, which see a free publicity in getting a good position into the public rankings (publish in newspapers, magazines, etc.).

Several papers have examined if the categories created by grouping the mutual funds according to their attributes are the same as those assigned by the investment objectives. The seminal work, developed by Sharpe (1992), determined the mutual fund styles applying an asset class factor model (twelve asset classes are employed). In such model each factor represents the return on an asset class and the sensitivities (estimated parameters) are required to sum 1. This is employed to provide a natural method for constructing benchmarks, through a portfolio compounded by a mix of asset classes with the same estimated style.

Brown and Goetzmann (1995) found the mutual fund managers who were losses during the first half of the year; tend to increase the volatility of their mutual funds in contrast to the winner mutual funds for the same time period. The goal of this behaviour is to offer a better annual ranking between the mutual funds in their group. However, the level of risk is absolutely different. Also, Brown and Goetzmann (1997) propose a new classification method to solve the problem of determining mutual fund styles. This method captures nonlinear patterns of returns that result from active portfolio management styles and is insensible to strategic “gaming” of benchmark. The authors found that existing classifications do a poor job at forecasting differences in futures performances.

Di Bartolomeo and Witkowski (1997) do a classification of mutual funds based on the Sharpe’s methodology (1992), concluding that more than one to three mutual funds are misclassified. They apply Monte Carlo simulations to built randomly equally weighted portfolios from both, the official classification and their new classification method. They conclude that latter outperforms the official classification, in terms of Sharpe’s Ratio. From the Spanish case, Matallín and Fernández (1997) find that the obtained styles from the Sharpe’s asset class factor (1992) is in agreement with the official classification, which was established by CNMV[2].

Marathe and Shawky (1999) classify mutual funds employing cluster analysis and study whether these categories created by clusters are the same of those based on their investment objectives. They find that a 43% of the mutual funds do not belong to their stated categories and that in many instances self-declared categories of mutual funds are indistinguishable from one another when their classification is based on financial characteristics.

Kim et al (2000) classify mutual fund using observable attributes (characteristics and investment style) in addition to the risk and return measures. They implement a discriminant analysis. They find that the stated objectives of over 50% of the funds do not match their attributes-based objectives, and the stated objectives of over 33% of the

[2] It must be note that this official classification was changed in 1999 by the actual one. Our study is realized taking in account the latter, for this reason the results are not comparable with previous works.

funds depart severely from their stated objectives. These authors, like Brown and Goetzmann (1997) and diBartolomeo and Witkowski (1997), conclude that the current system of classifying mutual funds based on their stated objectives has a significant room for improvement. This provides impetus for research into better and more careful fund classification techniques to ensure that the funds do indeed stay true to their stated objectives. Then, according with the above idea, our paper applies a combination of nonlinear methods, at the object to improve the actual mutual fund classification.

The remainder of the paper is organized as follows. In the next section we describe the dataset used for clustering. In the third section we will briefly describe the models employed, K Nearest Neighbors (KNN), Self-Organizing Maps (SOM) and K-means. In the forth section we will detail the procedure that we follow, that is the method employed to implement the models used and to classify the mutual funds. In this section we also analyze and summarize the main results obtained. Section five concludes the paper.

2. Dataset

Our database consists of the 1592 mutual funds from the Spanish market, which contain weekly data from 1st January 2000 until 31st December 2002, representing approximately the seventy per cent of the whole mutual fund Spanish world (2333 funds). The funds are classified into 14 different categories. These categories, as we mentioned before, were established by the CNMV and INVERCO in June, 1999, and constitutes the actual and official classification. In this sense, this study supposes a big and novel contribution, given its relevance for particular and institutional investors. In Table 1, a survey of the 14 categories and their mainly requisites are shown3.

A whole description (economical and statistical) of the database is shown in Table 2. We show, in columns, for each category: the mean return for the last year (return), the mean return for the whole time span (return3), the risk for the last year (risk) and for the whole time span (risk3), the skewness, the kurtosis, the five per cent highest profits and losses for the last three years, the number of mutual funds in each category, the percentage that our data represents from total number of funds in Spanish market, and the last column represents the percentage of funds that rejects the null hypothesis of normality from a Jarque-Bera test, at ten percent of significativity.

We consider a total of ten fund attributes, which we group in four different sets. These attributes are: (1) average return, (2) standard deviation, (3) skewness, (4) kurtosis, (5) the five percent of maximum losses, (6) the five percent of highest weekly returns, (7) the Reward-to-Semivariability ratio 4, (8) the Beta from the IBEX-35, (9) the Beta from the Notional Bond's ten years duration and (10) the correlation of each fund with an equally weighted benchmark obtained from each one of the 14 official categories. Notice that all of the above attributes were computed both for the last year and for the whole time span (three years).

[3] As it can be seen some of the official categories do not have any special and defined requirements, like Global Mutual Funds. We decided to include these mutual funds in order to really know what its behavior is in a financial frame, given its relevance for particular investors.

[4] We consider this ratio instead of the Sharpe's Ratio because of the non normality found between the mutual funds. A common mistake in the literature is to consider Sharpe's Ratio when the returns distribution is non normal (we can see in Table 2 how in average more than 85 % of funds reject the null hypothesis of normality). As it was shown by Chen and Knez (1996) the Sharpe's Ratio has got a bias in this cases.

Contrarily to similar studies (e.g. Kim et al., 2000) we do not takes the net asset value of mutual funds as a relevant variable for clustering. We consider that it does not reflect any specific behavior of funds in a financial frame, given that the liquidity is always guaranteed in mutual fund universe.

3. Applied Techniques.

In this paper we present a study about mutual fund classification which is implemented in four steps on the basis of clustering techniques.

It should be emphasized that the clustering, here, is carried out using very computationally intensive methods that are:

- The **K Nearest Neighbors** (KNN) algorithm, and
- A two-level approach, where the data set is first clustered using the **SOM**. Next the obtained results from the net are clustered employing the **k-means** algorithm. Is important consider that this approach clusters the SOM rather than clustering the data directly.

3.1. The K Nearest Neighbors (KNN) algorithm

The k nearest neighbor (KNN) algorithm is one of the most honored methods in machine learning, that are very popular non parametric models[5].This algorithm store in

[5] See Dasarathy 1991 and Aha 1990 for a survey.

memory the whole training set and, and Euclidean distance (possibly weighed) is computed to match inputs to training samples and cluster them. Thus, to classify a new example, the algorithm proceeds as follow:

i) The above mentioned distance between the new example and each stored training example is calculated.

ii) Next, the new example is classified into the group of the k nearest neighboring example, that is, into this one from which the distance is minimum.

The strategy for classifying involves searching from the training cases and finding those whose inputs are closest to the inputs of the test case. Neighborhoods are defined in terms of the number of k neighbors to use.

3.2. The Self-Organizing Map

Artificial Neural Networks (ANN) can be thought about as non linear, multi-layered, parallel regression methods. We can distinguish between supervised and unsupervised ANN.

Following Deboeck and Kohonen (2000), supervised Neural Nets are techniques for extracting from data input-output relationships and for storing those relationships into mathematical equations that can be used for forecasting or decision-making. The ANNs learns trough an adaptive, iterative process to capture the relationship the given inputs and the outputs, thus is necessary that the user specify the desired outputs. After the training phase, the ANN is capable to generalize, this means that it can be used on data that is has never seen.

The SOM acronym stands for Self-Organizing Map, also called Kohonen map, a popular feed forward neural network based on unsupervised learning which has properties of both vector quantization (Gray, 1984) and vector projection algorithms (Kaski, 1997). The SOM, through a process called self-organization, configures the output units into a topological representation of the original data, positioning the prototype vectors on a regular low-dimensional grid in an ordered fashion, making the SOM a powerful visualization tool.

The number of neurons may vary from a few dozen up to several thousand. They are connected to adjacent neurons by a neighborhood relation, dictating the topology of the map. Each neuron *i* has an associated *d*-dimensional prototype or codebook vector, the weigh vector $\boldsymbol{m_i} = [\boldsymbol{m_{i1}}, \boldsymbol{m_{i2}}, \ldots, \boldsymbol{m_{id}}]$. The dimension *d* is the same than the input vector's dimension.

Each neuron has actually two positions: one in the input space (the prototype vector) and another in the output space (the map grid). Thus, SOM is a projection method defining a nonlinear projection from the input space to a lower-dimensional output space. Also, during the training the prototype vectors move so that they follow the probability density of the input data. Thus, SOM is a vector quantization algorithm too.

The SOM is trained iteratively. In each training step, one sample vector $\boldsymbol{x}$ from the input data set is chosen randomly and the similarity between it and all the weigh vectors of the map are calculated using some distance measure, typically Euclidian distance. The unit whose incoming connection weights have the greater similarity with the input pattern $\boldsymbol{x}$ is called the Best-Matching Unit (BMU), denote as c:

$$\|x - m_c\| = \min_{i}\{\|x - m_i\|\} \tag{1}$$

Where $\|\bullet\|$ is the distance measure.

After finding the BMU, the connection weights (prototype vectors) of the SOM are adjusted. SOM creates a topological mapping by updating not only the BMU's weights, which are adjusted (i.e. moved in the direction of the input pattern by a factor determined by the learning rate[6]), but also adjusting the weights of the adjacent output units in close proximity of in the neighbourhood of the winner. So not only does the BMU get updated, but the whole neighbourhood of output neurons gets moved closer to the input pattern.

The SOM update rule for the weight vector of neuron i is:

$$m_i(t+1) = m_i(t) + \alpha(t)h_{ci}(r(t))[x(t) - m_i(t)] \tag{2}$$

Where t denotes time, $\alpha(t)$ is the learning rate, that is a monotonically decreasing function of time between (0,1), and $h_{ci}(r(t))$ the neighbourhood kernel around the BMU

[6] This is the basic nature of competitive neural networks.

c, with neighbourhood radius $r(t)$, which typically decreases with time. The neighbourhood kernel is a no increasing function of time and of the distance of neuron i from the winner unit c.

$$h_{ci}(r(t)) = \exp\left(- \frac{\left\| r_c - r_i \right\|^2}{2\sigma^2(t)} \right) \tag{3}$$

Were r_c and r_i are positions of units c and i on the SOM grid.

The SOM training algorithm that we implement is the Batch training algorithm. This is an iterative algorithm in which, opposite to the traditional sequential training, in each training step the data set is shown to the SOM as a whole, and the new weight vectors are weighted averages of the data vectors over the samples, where the weighting factors are the neighbourhood function values. That is, the data set is partitioned according to the Voronoi regions of the map weight vectors. After this, the new weight vectors are calculated as follow:

$$m_i(t+1) = \frac{\sum_{j=1}^{n} h_{ic(j)}(t) x_j}{\sum_{j=1}^{n} h_{ic(j)}(t)} \tag{4}$$

Where $c(j)$ is the BMU of sample vector x_j, $h_{i,c(j)}$ the weighting factor (neighbourhood function), and n is the number of sample vectors.

The training is done in two phases:

- Rough training: with large (initial) neighbourhood radius and large (initial) learning rate.
- Fine-tuning training: with small radius and learning rate.

For a more complete description of the SOM please refers to Kohonen (1997).

3.3. The K-means algorithm

The k-means constitute one the most popular methods for multidimensional data clustering. It is an iterative relocation algorithm, which works in a partitioning way. This algorithm minimizes the within group sum of squares between all points and the luster center (*centroid*) to get the best classification of the data.

Following Tou and Gonzalez (1974), the procedure in which this algorithm works can be defined as follows:

i) K initial centers $z_1(1)$, $z_2(1)$, …, $z_k(1)$ are choose. They may or may not be the centroids.

ii) All the data points are assigned to clusters. At the k-th iterative step, each one of the whole sample is assigned among the K clusters according to the following relation

$$x \in C_j(k) \text{ if } \|x - z_j(k)\| < \|x - z_i(k)\| \tag{5}$$

for all i=1, 2, …, K; $i \neq j$; where $C_j(k)$ denotes the set of samples whose centroid is $z_j(k)$.

iii) Next, the new centroids are computed [z_j (*k+1*), j = 1, 2… K] such as the sum of the squared distances from all points in C_j (*k*) to the new cluster center is minimized. Knowing that the data mean is the measure which minimizes this distance, then, the new cluster center is:

$$z_j(k+1) = \frac{1}{N_j}\left(\sum_{x \in C_j} x\right) \qquad j = 1, 2, \ldots, K \qquad (6)$$

where N_j is the number of data in the new cluster C_j (*k+1*).

iv) When there not exists any improvement, that means $z_j(k+1)=z_j(k)$ for all j, then the algorithm has converged and the procedure is finished. Whether this does not happen, all steps from number ii must be realized again.

As it can be noted, the configuration of the final clustering will depend on the initial centroids and on the number of K. To solve this, we employ a common measure: The Davies-Bouldin Index (1979). This index has the objective to minimize the within-cluster distance and maximize the between-cluster separation[7]. This index can be defined as:

$$\frac{1}{C}\sum_{k=1}^{C} \max_{k \neq p}\left\{\frac{S_c(Q_k) + S_c(Q_p)}{d_{ce}(Q_k, Q_p)}\right\} \qquad (6)$$

where S_c is the within-cluster distance and d_{ce} is the between clusters distance. C is the number of clusters

The K-means algorithm, as any other partitive methods, makes implicit assumptions on the form of clusters. Therefore, k-means tries to find spherical clusters. Besides, It must be noted that the k-means algorithm is equivalent to the SOM algorithm which neighbourhood radius *r(t)* is set to zero.

4. Methodology and results

We structure our results in five steps.

- **Step 1: Evaluation of the stated categories**

Firstly, we evaluate the 14 official mutual funds categories according to the CNMV. We employ the knn algorithm and an out-of-sample technique similar to jackknife[8] to evaluate the percentage of misclassified mutual funds.

The methodology of this first part is as follows:

i) We train the knn algorithm for the whole sample (for a given set of attributes) except for one fund (*mutual fund i*).

ii) Next, we present this mutual fund (with a group of attributes) to the knn and it is clustered in some of the 14 official groups. Then, we compute if it is correctly classified or not. Thus, if it is misclassified, we compute where the knn places it.

---

[7] Although, several others measures have been proposed to achieve the same goal (see Bezdek and Pal, 1998) we decided to employed this because it is followed by Vesanto and Alhoniemi (2000) as a complementary tool to Self-Organizing Maps (SOMs).

[8] For more details see Efron and Tibshirani (1998).

iii) After that, we repeat the steps i and ii for another fund (*mutual fund i+1*), and so for all mutual funds.

As it can be seen it is a very intensive computational process, because we have to train the knn algorithm a number of times equal to the number of mutual funds in our database (1592 times). Moreover, we not only do it for a set of attributes, we take four groups of attributes and we repeat it for each one. The group of attributes taken is:

a) Group I: The return and risk for the last year and for the entire sample (4 attributes).
b) Group II: The attributes of Group I, the Skewness and the Kurtosis for the last year and the entire sample (8 attributes).
c) Group III: Those in Group I and the correlation of each mutual fund with an equally weighted benchmark from each category (32 attributes).
d) Group IV: The attributes of Group I, the five percent of maximum losses, the five percent of highest returns, the reward-to-semivariability, the beta from the IBEX-35 index and the beta from a Notional Bond index. All of them are computed both for the last three years and for the last one. (14 attributes)

These groups are maintained during each one of the four steps in which we divide our analysis.

We show the results for these experiments in Tables 3.1-6.5. As it can be seen the average mutual funds misclassified rate is about 40%, which is according with previous studies from the E.E.U.U (e.g. Kim et al. 2000). We can also note that the misclassified rate does not depend on the considered number of neighbours. The results for a group of attributes are very similar consistent for 1, 5, 10, 15 and 25 neighbours.

We can observe how the first official group (FIAMM) is the one with the lowest rate of funds misclassified. It is classified correctly over 80%. However, categories 2 (GLOBAL) and 9 (RV NACIONAL) are classified like mutual funds in some other categories over 90%. It could has a negative effect on the performance of a portfolio diversified between mutual funds categories, because an investor could try to diversified between official categories and they are not really a different ones.

Therefore, from this first analysis we can conclude that the stated objectives are not according with the actual objectives the funds pursue. This means that the official classification is not according with their attributes (characteristics, risk/return measures and performance measures). Therefore, we realize another classification following a natural clustering technique.

- **Step 2: Building some new categories**

In this new step, we applied the above mentioned two-level approach that, based on both the SOMs and the clustering algorithm k-means, tries to find a natural classification of Spanish mutual funds according to our four groups of attributes.

The SOM is especially suitable for data survey because it has prominent visualization properties. It creates a set of prototype vectors representing the data set and carries out a topology preserving projection of the prototypes from the d-dimensional input space into a lower dimension space. But the problem is the visualization, and definition of the clusters from the SOM (For example, in Figure 1 we show the U-matrix for a SOM with 10x10 neurons grid and for the Group I of attributes).

As it can be seen it is quite difficult and subjective to definite the clusters from the U-matrix (matrix's distance between prototypes or neurons). In this paper, we consider a more objective approach based in applying k-means on the SOM prototypes (see Vesanto and Alhoniemi (2000) for a more detailed description of the methodology).

Firstly, we compute self-organizing maps of different sizes (5x5, 10x10, 10x15, 15x15 and 20x20 neurons) for each one of the four groups of attributes. All SOMs were trained employing the batch algorithm and were linearly initialized. The training was performed in two phases: a rough training phase with a higher learning rate and neighborhood width (we employ 500 epochs for this phase). And a fine-tuning phase with smaller learning rate and initial neighborhood width (we use 10.000 epochs to make this phase).

Next, we take the SOM's output and configure the clusters employing the k-means algorithm. To decide the right number of clusters, we compute the k-means from

20 to 2 clusters, and calculate the Davies-Bouldin (1979) Index. From these, we choose that number of clusters that minimize the Davies-Bouldin index[9].

We show the results of this process in Tables 7.1-7.4. As it can be seen there exists a different number of clusters from each attributes set. So for Group I of attributes we can divide the whole mutual funds into 3, 8 or 4 clusters. For all experiments, we decide taking the higher number of cluster or that number of cluster which is better distributed (do not have a huge amount of mutual funds in a unique cluster). Therefore, we choose as the best SOM clustering configuration: i) a 10x10 SOM for the Group I of attributes; ii) a 10x15 configuration for the Group II; iii) a 15x15 for the Group III; and iv) 10x10 neurons for Group IV. From here and for the rest of the paper, to simplify we only employ these SOM's configurations.

As we can see in Tables 7.1-7.4 the natural number of clusters for the Spanish mutual funds industry is smaller than the 14 categories establish by the CNMV. The number seems to be between 4 and 9, but not 14.

- **Step 3: Validation of the new classification**

We apply the same process than in step one to compute, from an out-of-sample perspective, the number of funds misclassified from the previous natural classification. The object is to see if this classification outperforms the official one. The results are shown in Tables 8.1-8.4.

[9] As the k-means initialization is random, we decided to repeat the above process 50 times for each SOM, to be sure that the number of clusters is the best one.

As it can be seen, the average rate of misclassifications is lower than from the 14 official categories. Therefore, we can conclude that there exists an erroneous stated objective classification and that it is possible to improve it. However, the particular investors are not interested in which is the misclassification rate of a classification. They will be interested in which one is better from a financial perspective. That means, which classification is better to achieve a diversified portfolio.

- **Step 4: Diversification effect**

In the forth part of the paper, we adopt the investors perspective and we suppose that they decide to make an equally weighted portfolio from each one of the clusters of official categories[10]. The process is as follows: We chose randomly a mutual fund from each one of the 14 official categories to make a portfolio, so this portfolio will have 14 assets. And we do the same for the obtained clusters from the two-level approach (step 2). Note that here we chose only a mutual fund from each cluster, so the number of assets in these portfolios is lower than in the official classification. This random process is repeated 10.000 times.

The results are shown in Table 9. As it can be seen the new classification from SOM and the k-mean algorithm, even the number of asset in each portfolio is lower, has achieved a better results in terms of return-risk (measured by the Sharpe's Ratio). Only in the case of doing the classification with the first group of attributes (risk and returns measures) the official category is not worse. However, we must realize that in the case of the Group I the total number of clusters is rather lower. It has only four clusters, and therefore, only four assets in its portfolio.

[10] Experiments like that were done by some other authors: Gallo and Lockwood (1997) or diBartolomeo and Witkowski (1997).

Therefore, we can conclude that the Spanish official classification according with the mutual funds stated objectives is wrong. It has an excessive number of categories whose effect is a reduction in the diversification effect that an investor can get with it.

- **Step 5: A hybrid classification SOM-CNMV**

In this last part of the paper, we present a new classification, which is a mix between official one and that from the SOMs. We allocate each category from the official classification to that cluster (obtained in the second section applying the two-level approach) in which the highest percentage of this category was classified[11].

For example, if we look at the Table 7.1, the second net (which has a 10x10 topology) we will place the Category I from the CNMV in cluster 1, the Category II in cluster 7, and so on. Next, we compute the knn and jackknife technique to know if this "hybrid" classification has a higher or lower misclassification rate.

We show the results in Tables 8.1-8.3. And we can conclude that, even, for this mixed classification the rate of misclassified is reduced. In this case, the average rate of mutual funds correctly classified is about 80% (except for Group II of attributes), which is about 10% higher than from the stated objective classification. However, it must be note that the number of clusters has been reduced. It could indicate that there are not such different types of mutual funds investment strategies as the CNMV publish.

[11] It must be note that it is not needed to have a percentage over 50%, it depends on the number of clusters.

5. Conclusions

In this paper, we analyze whether Spanish mutual funds are misclassified according to the actual official classification from CNMV, which is based in the investment objectives. We find that only an average of 60% of the mutual funds is correctly classified according to their financial attributes (risk/return, correlations between groups and performance measures). We also found that this erroneous classification is debt to an excess of categories. Some of the categories based in the investment objectives are not really different enough from others in financial terms; therefore, they should be eliminated.

This fact could be debt to a gaming process. This point out that funds manager declares some investment objective at the beginning of the mutual fund. But, during the mutual fund's life they escape from that restriction. They do it trying to get a higher position in the public rankings.

We investigate whether if it is possible to make a new classification according to the mutual fund attributes, which has a lower rate of misclassified funds. To do it, we employ a two-level approach, which is based on Self-Organizing Maps (a type of unsupervised Neural Nets) and k-means (a very popular clustering algorithm). We find that it achieves a lower rate of misclassified in an out-of-sample experiment. Besides, we find that the number of optimal categories must be between four and nine, but never fourteen.

From an investor perspective, the most relevant is to have a classification which permits to diversify and to achieve a good performance in term of risk/return. We do an application to portfolio theory, simulating equally weighted portfolios. Each portfolio is compounded by only one mutual fund from each category. We find that the new classification outperforms the official one, in term of Sharpe's Ratio. The portfolios based on official classification are worse, although it has 14 mutual funds and the new classification has between 4 and 8.

Moreover, we propose another classification which is a hybrid between the official one and the classification obtained from the SOMs and k-means. We reallocate each one of the official categories in the cluster, obtained form SOMs, that this category has a higher proportion of elements. Again, this new classification outperforms the official one.

References

Aha, D,W., 1990. A Study of Instance-Based Algorithms for Supervised Learning Tasks. Technical Report, University of California, Irvine.

Bezdek, J.C.; Pal, N. R., 1998. Some New Indexes of Cluster Validity. *IEEE Trans. Syst. Man. Cybern,* vol. 28, 301-315.

Brown, S.J.; Goetzmann, W.N., 1995. Performance Persistence. *Journal of Finance*, v50(2), 679-698.

Brown, S.J.; Goetzmann, W.N., 1997. Mutual Fund Styles. *Journal of Financial Economics*, 3, vol. 43, 373-399.

Chen, Z.; Knez, P., 1996. Portfolio Performance Measurement: Theory and Applications. *The Review of Financial Studies*, Summer, 9(2), 511-555.

Dasarathy, B.V., 1991. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press.

Davies, D. L.; Bouldin, D.W., 1979. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Machine Intell*, vol.1, 224-227.

Deboeck, G.; Kohonen, T., 2000. Visual Explorations in Finance and Investments Using Self-Organizing Maps. Springer-Verlag, Berlín.

diBartolomeo, D.; Witkowski, E., 1997. Mutual Fund Misclassification: Evidence based on Style Analysis. *Financial Analysts Journal*, Sep/Oct, 32-43.

Efron, B.; Tibshirani, R., 1998. An Introduction to the Bootstrap. Chapman&Hall-CRP, cop.

Gallo, J.; Lockwood, L.J., 1997. Benefits of Proper Style Classification of Equity Managers. *Journal of Portfolio Management*, vol. 23, 3, Spring, 47-55.

Gray, R.M., 1984. Vector Quantization. *IEEE ASSP Magazine*, April, 4-29.

Grinblatt, M.; Titman, S., 1994. A Study Of Monthly Mutual Fund Returns And Performance Evaluation Techniques. *Journal of Financial and Quantitative Analysis*, v29(3), 419-444.

Kaski, S., 1997. Data Exploration Using Self-Organizing Maps. Ph. Thesis, Helsinki University of Technology. Acta Polytechnica Scandinavica: Mathematics, Computing and Management in Engineering.

Kim, M., Shukla, R.; Thomas, M., 2000. Mutual Fund Objective Misclassification. *Journal of Economics and Business*, v52 (4,Jul/Aug), 309-324.

Kohonen, T., 1997. Self-Organizing Maps. Springer, Berlín.

Marathe, A.; Shawky, H.A., 1999. Categorizing Mutual Funds Using Clusters. *Advances in Quantitative Analysis of Finance and Accounting*, 1, vol. 7, 199-204.

Matallín, J.C.; Fernández, M.A., 1999. Análisis de la Performance a través del Estilo del Fondo de Inversión. *Revista Española de Financiación y Contabilidad*, 99, 413-442.

Sharpe, W.F., 1992. Asses Allocation: Management Style And Performance Measurement. *Journal of Portfolio Management*, v18(2), 7-19.

Tou, J.T.; González, R.C., 1974. Pattern Recognition Principles. Massachusetts: Addison-Wesley.

Vesanto, J.; Alhoniemi, E., 2000. Clustering of the Self-Organizing Map *IEEE Transactions on Neural Networks*, vol.11, 3, May, 586-600.