International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Experiments on Clustering Algorithms for Mixed and Incomplete Data

Yamilé Hernández Echemendía

Abstract*: Clustering mixed and incomplete data is a goal of frequent approaches in the last years because its common apparition in soft sciences problems. However, there is a lack of studies evaluating the performance of clustering algorithms for such kind of data. In this paper we present an experimental study about performance of seven clustering algorithms which used one of these techniques: partition, hierarchal or metaheuristic. All the methods ran over 15 databases from UCI Machine Learning Repository, having mixed and incomplete data descriptions. In external cluster validation using the indices Entropy and V-Measure, the algorithms that use the last technique showed the best results. Thus, we recommend metaheuristic based clustering algorithms for clustering data having mixed and incomplete descriptions.

***Keywords* : cluster validation, data clustering, incomplete data, mixed data.**

I. INTRODUCTION

The separation of objects into groups, so that the objects in the same group have a high level of similarity, and in turn, that do not resemble the objects of other groups is known as Unsupervised Classification, or Data Clustering [1, 2]. Three types of disciplines are dedicated to this type of tasks: Pattern Recognition, Artificial Intelligence and Data Mining. From different points of view, but with many elements in common, these disciplines aim to solve the problem of automating complex tasks, and processing information volumes beyond human capabilities in several domains [3-9].

In general, the clustering algorithms can be classified into three large groups according to the type of data they handle: algorithms designed for the management of numerical data, those designed for the management of categorical data, and those designed for the handling of mixed data [10, 11].

In the first two cases, algorithms of proven efficacy exist in the literature, and several of them are part of systems for commercial purposes. However, in the latter case, the amount of algorithms proposed is significantly smaller, and it is not until the present century that this problem has been tackled with some force in the scientific community [12, 13]. This is because worldwide a large amount of data is collected from numerous areas, and many of these data are described simultaneously by numerical and categorical attributes, and sometimes there may be absences of information regarding the value of a given attribute in a data (object). Problems with mixed and incomplete data (DMI) are commonly found in so-called non-formalized sciences (soft sciences) such as medicine, geosciences, criminalistics and others [14-20].

The solution to problems with DMI has been addressed in the literature using four variants [21]:

- Convert qualitative data into numerical. Each value of each feature is assigned a numerical encoder. This has the disadvantage that the qualitative data are then added, multiplied and averaged, losing the initial sense of the data, since the codes are not numbers [22, 23].
- Convert numerical data into qualitative. For this, discretization techniques are generally used. There are cases of variables that cannot be discretized due to their nature, an example of this is the Bouguer anomaly, in the forecast of the maximum magnitudes of earthquakes in the Caribbean region.
- Analyze separately quantitative data and qualitative data. In this variant, possible semantic dependencies between data of different nature are not taken into account. When separating the numerical variables from the rest, this type of situation is not analyzed.
- Work directly with DMIs using heterogeneous dissimilarities for numerical and non-numerical data. In this case, no transformations are required, so no information is lost. This variant is considered to be the most semantic and closest to the way natural and social science specialists work in reality [24-26].

In most applications, researcher use the first three variants, despite their drawbacks [27-30]. However, we are interested in the later variant, that is, working with DMI.

This paper analyzes the behavior of different clustering algorithms that use the last approach by directly manipulating DMI. This way we can determine which methods and approaches that perform better in the DMI cluster. All algorithms operate on a database $O=\{o_1, \ldots, o_n\}$ of objects. The descriptions $A(o)$ of each object have the same attribute number, $A=\{a_1, \ldots, a_d\}$, where $A_i(o_j)$ denotes the value of the ith attribute of the jth object. In addition, the existence of a dissimilarity function D capable of handling the descriptions of the objects in terms of A is assumed [31-34]. In this work, only the algorithms capable of obtaining partitions are considered $C=\{c_1, \ldots, c_k\}$ of the objects of O, in this way no object belongs to more than one group. On the other hand, according to the number of groups to obtain the clustering algorithms can be divided into two broad categories: the free unsupervised classification algorithms (where the desired number of groups is not defined a priori), and the algorithms of restricted unsupervised classification (where the number of groups to obtain is defined a priori). In this work, only restricted clustering algorithms are treated.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
Yamilé Hernández Echemendía*, Department of Computer Science, University "Máximo Gómez Báez", Ciego de Ávila, Cuba. Email: wnicole110@gmail.com

From here the rest of the work is organized as follows: In section 2 we present the grouping methods used in the comparison. In section 3 we show the quality measures used for group validation. In section 4 we present an analysis of the experiments performed on different data sets and the results obtained. Finally, the conclusions are given in section 5 possible.

II. CLUSTERING MIXED AND INCOMPLETE DATA

As mentioned earlier, the problem of unsupervised classification of mixed data has received less attention in the international scientific community than its counterparts for purely numerical or categorical data [2], and to our knowledge the first works in this regard date back to the end of the decade from the 90 of the last century. In this section, several of the algorithms reported in the literature for restricted DMI clustering are analyzed. It should be noted that only the algorithms that handle the mixed data as such are analyzed, but not those that encode the categorical features or discretize the numerical features.

The existing grouping algorithms for DMI, for the most part, are the product of extensions made to methods for handling homogeneous data types (numerical or non-numerical), these can be divided into several categories according to the procedure they use to group objects.

A. Partition clustering

The strategy used by the partition algorithms to find the groups is to iteratively relocate objects between subsets. In 1967 J. MacQueen proposed a classic algorithm that has been called k-means which is considered archetype of the partition model [35]. The k-means algorithm for its operation has to know a number k, which is the number of groups that you want to obtain. The idea is to locate k centroids and group the objects by their nearest centroid according to a distance function defined a priori. Iteratively, the centroids are updated as the average of the objects that belong to each group, until the centroids stop changing.

Despite their low temporal complexity, O (I*k*m*n), where I is the number of iterations required for convergence, these algorithms have many disadvantages: For example, the final results depend on the initialization of the centroids. They are also invalid against objects with categorical attributes due to the need to calculate the average and are inadequate to detect non-convex groups and outliers (noisy objects). One of the first extensions made to k-means to deal with DMI was k-Prototypes published by Huang in 1997 [36]. This author bases his proposal on the definition of a new function is dissimilarity between objects, in such a way that it is possible to treat the mixed descriptions, it is presented below:

$$d(O_j, c_i) = \sum_{p \in Rn} (X_p(O_j) - X_p(c_i))^2 + \gamma_i \sum_{\substack{r \in Rc \\ O \in MI \\ O \neq O_{i_p}}} \Gamma(X_r(O_j), X_r(c_i)) \quad (1)$$

where: γ_i is a parameter, Rn is the set of numerical attributes, Rc is the set of categorical attributes, and $\Gamma(X_r(O_j), X_r(c_i))$ id equal to zero if $X_r(O_j) = X_r(c_i)$ and equal to one otherwise. In addition, it makes a modification in the way to obtain the centers of the groups taking as values the average of the numerical attributes, and the mode of the categorical attributes.

More recent in 2011, Ahmad and Dey proposed another modification to the classic k-means [37]. The modifications made consist of updating the dissimilarity function, taking into account the contribution of each attribute to each group. This algorithm does not have a name, only a pseudocode is stated under the title *modified_kmean_subspace_clustering*, so we decided for the purposes of this work, to refer to the algorithm as AD2011.

B. Hierarchical clustering

Hierarchical algorithms, as the name implies, build a hierarchy of clusters, joining or dividing the groups according to a certain similarity / dissimilarity function between the groups. In other words, they build a group tree called dendogram. Such an approach allows studying data with different levels of granularity. Hierarchical clustering algorithms are categorized as agglomerative (bottom-up) and divisive (top-down). An agglomerative grouping generally begins with unit groups (singleton clusters) and recursively joins two or more appropriate groups. The process continues until some stop criterion is reached (frequently the number k of groups). Among the advantages of hierarchical clustering algorithms, flexibility with respect to the level of granularity can be mentioned, they are easy to handle and are applicable to any type of attribute. Among the disadvantages are the non-improvement of the groups that have been constructed due to the non-consideration of the already assigned objects and the sensitivity to noise. In addition, the computational cost for most of these algorithms is also at least O (m2), where m is the number of objects, which limits their application to large data sets. In 1999, Reyes-González and Ruiz-Shulcloper proposed a new agglomerative algorithm for DMI [38]. The algorithm (which we will call AERE) in each step forms a new level, until all the objects are at the same level. A level is defined by the number of groups present in the level, the value of β0 (maximum similarity between two groups of the same level), and the set of possible partitions. The distinctive element is that it is deterministic (it always obtains the same solution), and that each group is made up of elements that are in the same connected β0 component in a graph of maximum similarity. Finally, the algorithm returns all possible structuring (set of partitions) for the desired level k of the hierarchy formed.Also for the DMI grouping, in 2005, a hierarchical algorithm called HIMIC (HIerarchical MIxed type data Clustering algorithm) was proposed [39]. The algorithm is based on the use of a dissimilarity function defined by the authors which is given by the following equation:

$$d(C_i, C_j) = \sum_{p=1}^{n} S_p(C_i, C_j) \quad (2)$$

For numerical attributes, $S_p(C_i, C_j)$ is given by equation (3), while for numerical attributes $S_p(C_i, C_j)$ is given by equation (4).

$$S_p(C_i, C_j) = 1 - \left| \frac{1}{|C_i|} \sum_{o \in C_i} X_p(o) - \frac{1}{|C_j|} \sum_{o \in C_j} X_p(o) \right| \quad (3)$$

$$S_p(C_i, C_j) = \sum_{l=1}^{|D_p|} a * b$$

$$a = \frac{\left| \{ o \in C_i \mid X_p(o) = d_l \} \right|}{|C_i|} \quad (4)$$

$$b = \frac{\left| \{ o \in C_i \mid X_p(o) = d_l \} \right|}{|C_i|}$$

Subsequently a traditional agglomerative method is applied, using as a stop criterion the obtaining of the desired number of groups k.

C. Optimization based clustering

The problem of grouping data can be seen as an optimization problem that locates the optimal centroids of the groups or finds the optimal partition of a set of objects. For this reason, different optimization techniques have been used successfully to help find the best solution or at least one solution good enough for a search space problem. In this sense, metaheuristics stand out as approximate algorithms that try to solve these problems, sacrificing the guarantee of finding the optimum in exchange for finding a "good solution in a reasonable time", which is why they have been used for the grouping of large groups of data.

It can be said that a metaheuristic is a high-level strategy that uses different strategies to explore the search space and because of its easy adaptation, simplicity and efficiency are among the most widely used approximate methods. Next, we will describe three proposals for DMI clustering algorithms that are based on metaheuristics.

The AKGA algorithm proposed in [40] It consists in the use of a Genetic Algorithm (GA) to obtain groups without the need to exhaustively apply a clustering algorithm. The application of a GA allows to leave local optimum, and search for a global optimum, without a computational cost too high. As a representation scheme, each solution consists of an individual, represented by a chain of size equal to the number of objects, and where each i-th element of the chain represents the group to which the ith object is assigned (see Fig. 1).

2	1	2	1	1	2	1	2	1	2

Fig. 1.Example of an individual. In this case, there are 10 objects, grouped into two groups. Each position encodes the group to which said object belongs

In the case of group centers, the authors use the Ahmad and Dey scheme of 2011 [37]. They also use the function of dissimilarity of objects to the centers of [37] as part of the genetic algorithm optimization function.

A metaheuristic that has been used successfully in the grouping of numerical data is ABC (Artificial Bee Colony) inspired by the natural behavior of bees and proposed by Karaboga for numerical optimization [41]. An extension to this to group DMI is BECA (BEe based Clustering Algorithm) [42]. This algorithm generates n initial clusters randomly that constitute the food sources. Then it generates new sources using the mutation strategy defined in [40] and using HEOM as dissimilarity [43] which allows dealing with DMI. For the evaluation of food sources use an internal validation index as an objective function, in this case the Dunn index [44] to obtain compact and well separated groups. Finally, after a number of iterations defined a priori, the food source (grouping) that optimized the objective function is returned.

Another metaheuristic used to group data is the so-called firefly optimization (Firefly Multimodal Optimization Algorithm, FMOA). This optimization technique was recently developed by Xin-She Yang [45] and be based on the social behavior of fireflies and the flashes of light they emit. DELUXE (Firefly flashes with life expectancy) [46], It is an algorithm developed to group DMI based on FMOA. Like BECA algorithm, its operation is based on the dissimilarity HEOM [43], however, unlike this in DELUXE, initially a number of fireflies is formed by selecting random centers and placing the objects around them according to [43].

Then an iterative process begins by comparing in pairs the attractions of the different fireflies (clusters) using the Silhouette index [44] as an objective function and fireflies with lower attractiveness are disturbed towards those of greater attractiveness. The disturbance consists in replacing one of the centers of the groups with another object randomly. To avoid falling into local optimum, a parameter was introduced to the algorithm: the life expectancy of the firefly, allowing replace the fireflies that could not be disturbed

III. CLUSTER VALIDITY

In the Supervised Classification the evaluation of the classification model is an integral part of its development process and there are measures and evaluation procedures of generalized acceptance, example: accuracy and cross-validation, respectively. However, data clustering is more complex to evaluate because when clustering algorithms are applied to a data set, the results differ [44]. On the other hand, depending on the particular problem addressed, a way of grouping (or structuring) may or may not be the most appropriate. In the specialized literature the concepts or methods for the quantitative and objective evaluation of the result of a clustering algorithm are called group validation indices. These quality functions are known as Cluster Validity Indexes. Validation indices are divided into two large groups: external and internal indices. External indexes, as the name implies, are focused on determining to what extent the grouping to be evaluated (AE) coincides with a model grouping (AM) of the data. These indices have a high practical utility. They allow to establish which (or which) of the groupings that were evaluated are closer to the correct grouping of the data.

Among the most commonly used external indexes is Entropy, for a detailed description refer to [47], which is based on the disposition of the AM groups with respect to the AE groups. Its equation is given by:

$$E = \sum_{k \in AE} \frac{|k|}{N} \left[\frac{1}{\log(|AM|)} \sum_{m \in AM} \frac{n_k^m}{|k|} \log \frac{n_k^m}{|k|} \right] \quad (5)$$

where: $|k|$ is the number of objects in the cluster k, $|AM|$ is the numer of clusters in the model grouping and n_k^m is the total number of objects in group k that belongs to the group of AM. The smaller the Entropy the better the grouping will be.

The *V-Measure* [47], It was proposed by Rosemberg and Hirschberg in 2007. This external index is based on Entropy, but takes into account two criteria: the homogeneity of a grouping is given if all groups contain elements of the same group in AM, and completeness if all Elements of a group in AM are in the same group of AE. It is given by the following equation:

$$V_\beta = \frac{(1+\beta) * h * c}{(\beta * h) + c} \quad (6)$$

Where h is homogeneity and c is completeness, then if β = 0, both completeness and homogeneity have the same weight.

On the contrary, if β>1, completeness is given with greater weight and if β<1, the greater weight is obtained by homogeneity. A maximum value of Measure V corresponds to a better quality grouping.

It is noteworthy that Measure V is considered as one of the most robust external indexes. For a detailed explanation of the calculation of h and c refer to [47].

IV. RESULTS AND DISCUSSION

In order to reach conclusions about the algorithms and approaches that have a better performance in the DMI cluster, the following methods were used: k-Prototypes (KP) [36], AERE [38], AGKA [40], HIMIC [39], AD2011 [37], BECA [42] and DELUXES [46].

The experiments were performed on a set of 15 labeled mixed databases (Table I) from the University of California in Irvine (UCI) repository [48].

The experimental comparison was all against all, running on a PC with Windows XP OS, 2.99 GB of RAM and Celeron (R) D processor at 3.06 GHz.

An important aspect in the design of the experiments are the parameters with which the algorithms will be executed. As all algorithms require knowing the number of groups to be formed, the value assigned to this parameter will match, for each database, the number of classes present. This allows class labels of each database to be taken as a model grouping; to then compare the latter with the grouping obtained by the algorithms [49].

In addition to the common parameters of the different algorithms they were given the same value. This allowed to achieve a certain homogeneity and reduce a possible imbalance in the performance of an algorithm against another by the use of different values for the same parameter.

In the case of dissimilarity experiments were performed using the HEOM function for all algorithms. The reason for its use was its good results in the treatment of IMD. The parameters used by each algorithm are shown in table II.

Due to the use of tagged databases, two external indexes were used as measures to assess the quality of the groups obtained by the algorithms.

Table- I: Characteristics of the databases used in the experiments

Database	Cat. Att.	Num. Att.	Classes	Missing	Objects
anneal	29	9	6	yes	798
autos	10	16	7	yes	205
cmc	7	2	3	no	1473
colic	15	7	2	yes	368
credit-a	9	6	2	yes	690
credit-g	13	7	2	no	1000
dermatology	1	33	6	yes	366
heart-c	7	6	5	yes	303
hepatitis	13	6	2	yes	155
labor	6	8	2	yes	57
lymph	15	3	4	no	148
postoperat-patient-data	7	1	3	yes	90
tae	2	3	3	no	151
vowel	3	9	11	no	990
zoo	16	1	7	no	101

Table- II: Parameters used by each algorithm

Algorithm	Parameters
CEBMDC	Similarity threshold: 0
AGKA	Number of generations: 10 Probability of mutation: 0.05 Population size: 10 Crossing probability: 1
AD2011	Group contribution parameter: 20
BECA	Number of food sources: 10 Number of scout bees: 1 Number of generations: 10 Limit of food sources: 10
DELUXES	Number of fireflies: 1 Life expectancy: 10 Number of iterations: 10

In this way, the groups for each database were obtained using the number of classes as the number of groups to be formed in each of the algorithms. The indices used were Measure V and Entropy analyzed in the previous section.

The process begins by applying the algorithms to each database and obtaining the respective groupings. These clusters are then evaluated using each index. Table III and Figures 2 and 3 show the results in terms of how many times each algorithm won, lost or tied with respect to the rest, with respect to Measure V, and with respect to it and Entropy, respectively. A triad of Wins, loses and ties (W-L-T) is shown in each cell, indicating the number of times: W means that measure V of the column algorithm was smaller than that of the row algorithm, L means the measure was greater and T means that there was a tie.

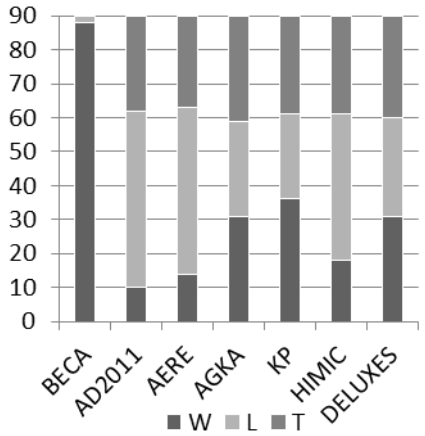

Fig. 2.Results of the performance of the methods using measure V as validation index

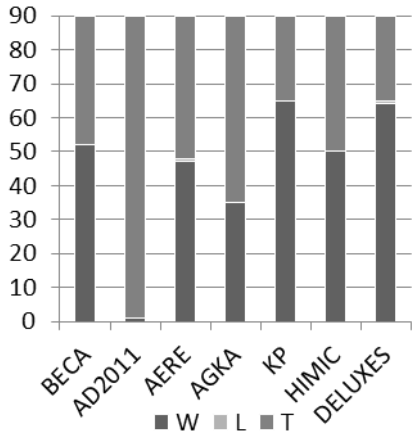

Fig. 3.Results of the performance of the methods using Entrpy as validation index

To compare the results of the algorithms, the Wilcoxon test was used for related samples [50].

Table- III: Win-Lose-Tie quantity for all algorithms taking into account Measure V

Algorithms	BECA	AD2011	AERE	AGKA	KP	HIMIC	DELUXES
BECA	-	0-15-0	1-14-0	0-15-0	0-15-0	0-15-0	1-14-0
AD2011	15-0-0	-	4-5-6	9-0-6	10-0-5	6-4-5	8-1-6
AERE	14-1-0	5-4-6	-	9-1-5	9-1-5	5-4-6	7-3-5
AGKA	15-0-0	0-9-6	1-9-5	-	5-3-7	3-6-6	4-4-7
KP	15-0-0	0-10-5	1-9-5	3-5-7	-	2-7-6	4-5-6
HIMIC	15-0-0	4-6-5	4-5-6	6-3-6	7-2-6	-	7-2-6
DELUXES	14-1-0	1-8-6	3-7-5	4-4-7	5-4-6	2-7-6	-
Total	**88-2-0**	10-52-28	14-49-27	**31-28-31**	**36-25-29**	18-43-29	**31-29-30**

Table- IV: Bilateral asymptotic significance of the Wilcoxon test for Measure V. The color of the cell indicates whether the value favors the row or column algorithm. won the row. won the column. there were no significant differences

Algoritmos	BECA	AD2011	AERE	AGKA	DELUXES	HIMIC	KP
BECA	—						
AD2011	0.001	—					
AERE	0.001	0.65	—				
AGKA	0.001	0.004	0.008	—			
DELUXES	0.001	0.003	0.006	0.363	—		
HIMIC	0.001	0.363	0.496	0.14	0.033	—	
KP	0.001	0.011	0.031	0.496	1	0.173	—

Table- V: Bilateral asymptotic significance of the Wilcoxon test for Entropy. The color of the cell indicates whether the value favors the row or column algorithm. won the row. won the column. there were no significant differences

Algoritmos	BECA	AD2011	AERE	AGKA	DELUXES	HIMIC	KP
BECA	—						
AD2011	0.001	—					0.001
AERE	0.57	0.001	—				0.57
AGKA	0.009	0.001	0.14	—			0.009
DELUXES	0.281	0.001	0.008	0.011	—		0.281
HIMIC	0.82	0.001	0.334	0.006	0.307	—	0.82
KP	0.57	0.001	0.211	0.008	0.008	0.363	0.57

Finally, for each index, the differences between the groups obtained is determined by the previous statistical tests.

The p-values for the Wilcoxon test are shown in Table IV. According to this, the BECA algorithm was the one with the best performance according to Measure V since it is the one with the highest number of results with asymptotic significance not greater than 0.05, for 95% acceptability. Followed by the DELUXES algorithm, three times winner.

A similar process was carried out to compare the performance of the algorithms, using it as an Entropy validation index. The results of the Wilcoxon test are shown in table V.

According to Table IV, the DELUXES algorithm was the one with the best performance according to Entropy, since it has the highest number of results with asymptotic significance not greater than 0.05, for a 95% acceptability, followed by the BECA, HIMIC and KP respectively. Given the tests performed for the two indices, it was possible to conclude that the best performing algorithms were the BECA and DELUXES, obtaining first and second place respectively for Measure V, and reciprocally for Entropy.

V. CONCLUSION

In this work, a review of mixed and incomplete data grouping algorithms (DMIs) that use different approaches was performed, specifically the partition, hierarchical and metaheuristic-based ones were analyzed. In experiments carried out to evaluate performance, it was obtained that the DELUXES and BECA metaheuristic algorithms are suitable for the grouping of DMI, surpassing other methods. Using the V Measure as a validation index, BECA was higher than the rest of the methods, DELUXES being second. On the other hand, using Entropy, DELUXES was superior and BECA obtained second place. The experiments made allow us to conclude that metaheuristic based clustering have better results than the evaluated partition and hierarchical clustering algorithms, for mixed and incomplete data.

REFERENCES

1. Join, A.K., y R. C. Dubes, *Algorithms for Clustering Data*. 1988, New Jersey, USA: Prentice Hall.
2. Barroso, E., Y. Villuendas, and C. Yanez, *Bio-inspired algorithms for improving mixed and incomplete data clustering.* IEEE Latin America Transactions, 2018. **16**(8): p. 2248-2253.
3. Godínez, I.R., I. López-Yánez, and C. Yánez-Márquez, *Classifying patterns in bioinformatics databases by using Alpha-Beta associative memories*, in *Biomedical Data and Applications*. 2009, Springer. p. 187-210.
4. Villuendas-Rey, Y., M.M. Garcia-Lorenzo, and R. Bello. *Support Rough Sets for decision-making*. in *Fourth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support*. 2013. Atlantis Press.
5. Cerón-Figueroa, S., et al., *Instance-based ontology matching for e-learning material using an associative pattern classifier*. Computers in Human Behavior, 2017. **69**: p. 218-225.
6. Ortiz-Ángeles, S., et al., *Electoral Preferences Prediction of the YouGov Social Network Users Based on Computational Intelligence Algorithms.* J. UCS, 2017. **23**(3): p. 304-326.
7. Antón-Vargas, J.A., et al., *Improving the performance of an associative classifier by Gamma rough sets based instance selection.* International Journal of Pattern Recognition and Artificial Intelligence, 2018. **32**(01): p. 1860009.
8. Serrano-Silva, Y.O., Y. Villuendas-Rey, and C. Yáñez-Márquez, *Automatic feature weighting for improving financial Decision Support Systems.* Decision Support Systems, 2018. **107**: p. 78-87.
9. González-Patiño, D., et al., *A Novel Bio-Inspired Method for Early Diagnosis of Breast Cancer through Mammographic Image Analysis.* Applied Sciences, 2019. **9**(21): p. 4492.
10. Villuendas-Rey, Y., M. García-Borroto, and J. Ruiz-Shulcloper. *Selecting features and objects for mixed and incomplete data*. in *Iberoamerican Congress on Pattern Recognition*. 2008. Springer.
11. Acevedo, M.E., C. Yáñez-Márquez, and M.A. Acevedo, *Associative models for storing and retrieving concept lattices.* Mathematical Problems in Engineering, 2010. **2010**.
12. Villuendas-Rey, Y., *Maximal similarity granular rough sets for mixed and incomplete information systems.* Soft Computing, 2019. **23**(13): p. 4617-4631.
13. Villuendas-Rey, Y., et al., *NACOD: A Naïve Associative Classifier for Online Data.* IEEE Access, 2019. **7**: p. 117761-117767.
14. García-Borroto, M., et al. *Using maximum similarity graphs to edit nearest neighbor classifiers*. in *Iberoamerican Congress on Pattern Recognition*. 2009. Springer.
15. Villuendas-Rey, Y., et al., *Simultaneous instance and feature selection for improving prediction in special education data.* Program, 2017. **51**(3): p. 278-297.
16. Villuendas-Rey, Y., et al., *The naïve associative classifier (NAC): a novel, simple, transparent, and accurate classification model evaluated on financial data.* Neurocomputing, 2017. **265**: p. 105-115.
17. Villuendas-Rey, Y., et al., *An Extension of the Gamma Associative Classifier for Dealing With Hybrid Data.* IEEE Access, 2019. **7**: p. 64198-64205.
18. Medina-Pérez, M.A., et al. *Selecting objects for ALVOT*. in *Iberoamerican Congress on Pattern Recognition*. 2006. Springer.
19. Villuendas-Rey, Y., et al. *Simultaneous features and objects selection for Mixed and Incomplete data*. in *Iberoamerican Congress on Pattern Recognition*. 2006. Springer.
20. García-Borroto, M., et al. *Finding small consistent subset for the nearest neighbor classifier based on support graphs*. in *Iberoamerican Congress on Pattern Recognition*. 2009. Springer.
21. Ruiz-Shulcloper, J., *Pattern Recognition with Mixed and Incomplete Data.* Pattern Recognition and Image Analysis, 2008. **18**(4): p. 563-576.
22. Cerón-Figueroa, S., et al., *Instance-based ontology matching for open and distance learning materials.* The International Review of Research in Open and Distributed Learning, 2017. **18**(1).
23. García-Floriano, A., et al., *Social Web Content Enhancement in a Distance Learning Environment: Intelligent Metadata Generation for Resources.* International Review of Research in Open and Distributed Learning, 2017. **18**(1): p. 161-176.
24. Hernández-Castaño, J.A., et al., *Experimental platform for intelligent computing (EPIC).* Computación y Sistemas, 2018. **22**(1): p. 245-253.
25. Villuendas-Rey, Y., et al., *Medical Diagnosis of Chronic Diseases Based on a Novel Computational Intelligence Algorithm.* J. UCS, 2018. **24**(6): p. 775-796.
26. Yáñez-Márquez, C., et al., *Theoretical Foundations for the Alpha-Beta Associative Memories: 10 Years of Derived Extensions, Models, and Applications.* Neural Processing Letters, 2018. **48**(2): p. 811-847.
27. López-Yáñez, I., L. Sheremetov, and C. Yáñez-Márquez, *A novel associative model for time series data mining.* Pattern Recognition Letters, 2014. **41**: p. 23-33.
28. Lytras, M.D., et al., *The Social Media in Academia and EducationResearch R-evolutions and a Paradox: Advanced Next Generation Social Learning Innovation.* J. UCS, 2014. **20**(15): p. 1987-1994.
29. López-Yáñez, I., et al., *Collaborative learning in postgraduate level courses.* Computers in Human Behavior, 2015. **51**: p. 938-944.
30. García-Floriano, A., et al., *Support vector regression for predicting software enhancement effort.* Information and Software Technology, 2018. **97**: p. 99-109.
31. Villuendas-Rey, Y., Y. Caballero-Mota, and M.M. García-Lorenzo. *Using rough sets and maximum similarity graphs for nearest prototype classification*. in *Iberoamerican Congress on Pattern Recognition*. 2012. Springer.
32. Villuendas-Rey, Y., Y. Caballero-Mota, and M.M. García-Lorenzo. *Intelligent feature and instance selection to improve nearest neighbor classifiers*. in *Mexican International Conference on Artificial Intelligence*. 2012. Springer.

33. Villuendas-Rey, Y., Y. Caballero-Mota, and M.M. García-Lorenzo. *Prototype selection with compact sets and extended rough sets*. in *Ibero-American Conference on Artificial Intelligence*. 2012. Springer.
34. Villuendas-Rey, Y., et al. *Nearest prototype classification of special school families based on hierarchical compact sets clustering*. in *Ibero-American Conference on Artificial Intelligence*. 2012. Springer.
35. MacQueen, J. *Some Methods for Classification and Analysis of Multivariate Observations*. in *Fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1967. Berkeley.
36. Huang, Z. *Clustering large data sets with numeric and categorical values*. in *1rst Pacific - Asia Conference on Knowledge discovery and Data Mining*. 1997.
37. Ahmad, A., Dey L., *A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical data*. Pattern Recognition Letters, 2011. **32**: p. 1062-1069.
38. Reyes-González, R., y Ruiz-Shulcloper, J. *Un algoritmo de estructuración restringida de espacios*. in *CIARP*. 1999. La Habana. Cuba.
39. Ahmed, R.A., et. al. *HIMIC: A Hierarchical Mixed Type Data Clustering Algorithm*. 2005. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.6369&rep=rep1&type=pdf.
40. Roy, D.K., Sharma, L. K., *Genetic k-means clustering algorithm for mixed numeric and categorical datasets.* International Journal of Artificial Intelligence & Applications (IJAIA), 2010. **1**(2): p. 23-28.
41. Karaboga, D., *An Idea Based On Honey Bee Swarm For Numerical Optimization*. 2005, Technical Report-TR06. Engineering Faculty. Computer Engineering Department, Erciyes University.
42. Cabrera-Venegas, J.F., *BECA algorithm for data clustering (In Spanish)*. 2012, University of Ciego de Avila.
43. Wilson, R.D., Martinez T. R., *Improved Heterogeneous Distance Functions.* Journal of Artificial Intelligence Research, 1997. **6**: p. 1-34.
44. Brun, M., et. al., *Model-based evaluation of clustering validation measures.* Pattern Recognition, 2007: p. 807-824.
45. Xin-She. Y., *Firefly algorithms for multimodal optimization.* Lecture Notes in Computer Sciences, 2009. **5792**: p. 169-178.
46. Hernadez-Echemendia, Y., *DELUX, DELUXE and DELUXES algorithm for mixed data (In Spanish)*. 2012, University Máximo Gómez.
47. Rosenberg, A., y Hirschberg, J. *V-Measure: A conditional entropy-based external cluster evaluation measure*. in *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2007. Prague.
48. Dua, D. and C. Graff. *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]*. 2019.
49. González-Patiño, D., Y. Villuendas-Rey, and A.J. Argüelles-Cruz, *The potential use of bioinspired algorithms applied in the segmentation of mammograms.* 2018.
50. Demsar, J., *Statistical comparison of classifiers over multiple data sets.* The Journal of Machine Learning Research, 2006. **7**: p. 1-30.

AUTHORS PROFILE

Yamilé Hernández Echemendía obtained her bachelor degree in Informatics from the University "Máximo Gómez Báez" of Ciego de Ávila, Cuba, in 2006. After working in the industry for five years, she obtained her M.S. degree in Applied Informatics from the same institution, in 2012. Currently, she is working as an adjoin professor at the Computer Science Department of the Computer Science Faculty of the University "Máximo Gómez Báez" of Ciego de Ávila, Cuba. Her research interests include firefly optimization, clustering, swarm intelligence and education. She is also interested in image segmentation applications for face detection and clustering applied to biometric identification.