Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 16, Iss. 1, pp. 26-40, 2018
© Brazilian Computational Intelligence Society

Clustering Students Based on Grammatical Errors for On-line Education

[1]Mariana G. da M. Macedo, [1,2]Elliackin M. N. Figueiredo, [1]Fabiana M. B. Soares, [1,2,3]Hugo Siqueira, [1]Alexandre M. A. Maciel, [2]Anuradha Gokhale and [1]Carmelo J. A. Bastos-Filho

[1]University of Pernambuco, Brazil
[2]Illinois State University, USA
[3]Federal University of Technology - Parana, Brazil
mgmm@ecomp.poli.br, emnf@ecomp.poli.br, fbms@ecomp.poli.br, hugosiqueira@utfpr.edu.br
amam@ecomp.poli.br, aagokha@ilstu.edu, carmelofilho@ieee.org

Abstract – Learning Management System (LMS) is an educational solution created for people who need flexibility regarding time and place. The problem of this kind of tool primarily concerns the difficulty in identifying which students have learned the content correctly. This paper aims to analyze the performance of a group of distance learning students regarding grammar errors in two different terms of an undergraduate course. Our hypothesis relies on the existence of different characteristics that emerge from subgroups of students with similar difficulties. This division can help tutors in educational platforms to develop specific recommendations tasks for each group of students. A previous work applied the well-known K-means algorithm to cluster the groups, but in that paper, we fixed the number of clusters. Therefore, we carried out a methodology to find the best number of clusters to be used in K-means for this problem. Moreover, we also applied the Fuzzy C-means to tackle the clustering problem and analyzed the results obtained by both algorithms using the well-known metrics in the literature (Gap Statistic and Davies-Bouldin) to assess the quality of the obtained groups. The experimental results showed that Fuzzy C-means approach outperforms the K-means algorithm. Moreover, the application of the Spearman Correlation on each group expose several differences, relations and similarities between groups and inside each one.

Keywords – Learning Management System, Clustering, Grammatical Errors, K-means, Fuzzy C-means, Gap Statistic, Davies-Bouldin, Spearman Correlation.

1 Introduction

The advances in information and communication technologies are providing the possibility to develop relevant educational methodologies and tools. Among them, we can cite the creation of Learning Management Systems (LMS) [1]. LMS developers keep improving the platforms' functionalities to help students to learn more effectively and efficiently [2].

Despite the flexibility provided by these platforms, it is hard to identify in the platform if the student rightly understood the given content. There are some reasons for this: firstly, the teacher may not perceive if students understand the content due to the lack of expected behaviors common in presential feedbacks. Secondly, these platforms manage and store a vast bulk of data, turning the real-time detection of undesired patterns not trivial. Finally, there is not enough feedback between tutors and students and vice-versa. A particular case involves checking the grammar accuracy on texts posted by students in discussion forums. Thus, it appears to be a good idea to have a technique which would identify useful patterns that could help the teachers or the LMS system itself to apply specific lessons for the students that are experiencing similar grammar difficulties to improve their learning [3].

In fact, an architecture using data mining tools for supporting students with grammar problems in an LMS context was proposed recently in [3]. According to this work, a generic structure can be used to provide recommendations of grammar contents and grammar exercises to be studied by each student in particular. This means that the system can present the capability to supply automatic recommendations of grammar lessons adjusted for each student. The hypothesis is that the students would make fewer grammar mistakes and thus write better after receiving this support. For applying these recommendations, it is necessary to arrange them according to similar grammar difficulties to make it possible to create specific materials for each group. This grouping process is essential since the number of students in the platform is huge. Thus, the analysis of each student is intractable. In this sense, the clustering task is fundamental to identify the groups of students with similar grammar errors.

In [3], the authors also proposed an approach for clustering students that are making the same type of grammar errors in *discussion forums*. The clustering of students was performed using a well-known and widely used algorithm called K-means. However, the work presented in [3] shows some drawbacks. Firstly, the number of clusters was established in advance based on a project decision, so the chosen number of clusters was three, which may be inappropriate for the problem. Thus, the first objective of this paper is to apply a methodological study similar to the one carried out in [4] to find the suitable number of clusters to be used by the clustering algorithm.

Moreover, the second objective of this paper is to investigate the performance of a more robust algorithm for clustering, called Fuzzy C-Means (FCM). In effect, because of the complexity of relations among the data and because of a possible overlapping on the data, the FCM may be a good alternative to K-means algorithm. A third contribution is to analyze the quality of the formed clusters of students using metrics and correlation techniques often used in the literature such as Gap Statistics, Davies-Bouldin, and Spearman's correlation.

The computational tools and metrics mentioned are well known in the literature, but from an educational point of view, their application on this problem can increase the capability of the platform to indicate the ways to help the professors in the students' learning process. This is still a challenge during the development of this kind of software [3].

this might bring to light more realistic information about students' difficulties. In addition, it can also help teachers to correlate mistakes when elucidating doubts or preparing grammatical Exercises.

Professors can provide more specific grammar lessons because of the grammar difficulties highlighted in each of the clusters. Thus, with the outcomes of the clustering process, professors can focus on students' difficulties and both professors and students can benefit from it.

In summary, the general objective of this paper is to create a tool that intends to support professors to assist students with grammar difficulties.

The remainder of this paper is organized as follows: Section 2 explains the necessary background to understand the proposal including the algorithms K-means and Fuzzy C-means and the metrics used to analyze the number of groups; Section 3 details the problem treated in this paper that consists of clustering students based on their grammatical errors; Section 4 exposes the experiments proposed to analyze the students similarities; Section 5 presents the conclusions.

2 Theoretical background

2.1 Clustering Algorithms

In this paper, we used K-means and C-means because of their simplicity and popularity. In Subsection 2.1.1 and 2.1.2, we detail the K-means and Fuzzy C-means algorithms, respectively.

2.1.1 K-means

K-means is a popular technique for clustering [5] [6] [7] [8]. This popularity is explained by its fast convergence and simplicity [9] [2]. The primary purpose of the K-means algorithm is to divide a bulk of data in a predefined number of groups. We perform this division considering the sum of each distance between the centroid and each instance from a particular group. The calculation of this distance can be done using Euclidean, City Block or Mahalanobis norms. In this paper, we use Euclidean distance because of the closeness between the values (their scale) [4]. The algorithm receives the data and the number of groups as inputs. The output is a set of groups, in which each group contains its respective data. The process is initialized with random centroids. Next, a loop is started to reassign instances and recalculate centers. This process is repeated over and over until a stop criterion is reached. The stop criterion can be the number of iteration or a pre-defined similarity. K-means minimizes the Sum of Squared Error (SSE) given by Equation 1: the number of groups is represented by K; the quantity of patterns within a cluster k is n_k; the centroid of each cluster is $\mathbf{c}_k$; $\mathbf{x}_i^k$ is i-th pattern within cluster k.

The centers $\mathbf{c}_k$ are update according to Equation 2, where the vectorial sum of all patterns in a given cluster is divided by the total number n_k of patterns within the cluster.

$$SSE = \sum_{k=1}^{K} \sum_{i=1}^{n_k} ||\mathbf{x}_i^k - \mathbf{c}_k||^2, \tag{1}$$

$$\mathbf{c}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{x}_i^k. \tag{2}$$

It is known that the main disadvantage of the K-means are its high dependence of the initialization, which can lead to a weak performance and increase the computational cost until the convergence. Besides, the algorithm do not deal well with overlapping data [9] [10].

2.1.2 Fuzzy C-means

Fuzzy C-means (FCM) overcomes the disadvantages of K-means, but continues to be a simple and fast algorithm [11] [10] [12]. FCM was used successfully in several applications [13] [14] [15]. Because of this, FCM appears as a good alternative to be applied in clustering task treated in this paper. Some of the advantages of FCM are: i) the determination of the number of groups can be performed automatically; ii) it works better with data overlapping than K-means; and iii) it is more robust to initialization, noise, and outliers.

The FCM algorithm does not require a predefined number of groups, but it is necessary a threshold to stop the convergence. As this threshold impacts the selection of groups' number, the automatic selection is influenced by the user.

Different from K-means, FCM does not create a hard partition of K clusters of a dataset $\mathbf{X}$ with size N, but defines a membership matrix $\mathbf{U}$ in which each entrance of the matrix is the membership matrix μ_{ij} of each pattern i for each cluster j. This membership matrix is defined by $\mathbf{U} = \{\mu_{ij}\}_{i,j}^{N,K}$ in which $\mu_{ij} \in [0,1]$, $\sum_{j=1}^{K} \mu_{ij} = 1, \forall i$, and $0 < \sum_{i=1}^{N} \mu_{ij} < N$.

The aim of FCM is to optimize the objective function given by Equation 3 in which $||.||$ is a norm, usually the Euclidean distance. To optimize the objective function, FCM applies iteratively two update processes using two different equations. FCM firstly updates the grade of membership μ_{ij} of each pattern i for each cluster j using Equation 4 where C is the maximum number of clusters. After that, FCM applies the update step to the centers using Equation 5, where m is the fuzziness coefficient provided by the user.

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{K} \mu_{ij}^{m} ||\mathbf{x}_i - \mathbf{c}_j||^2, \tag{3}$$

$$\mu_{ij}^{m} = \frac{1}{\sum_{k=1}^{C} \frac{||\mathbf{x}_i - \mathbf{c}_j||}{||\mathbf{x}_i - \mathbf{c}_k||}^{\frac{2}{m-1}}}, \tag{4}$$

$$\mathbf{c}_j = \frac{\sum_{i=1}^{N} \mu_{ij}^{m}.\mathbf{x}_i}{\sum_{i=1}^{N} \mu_{ij}^{m}}. \tag{5}$$

2.2 Metrics for performance evaluating

In this section, we briefly present the metrics deployed in this paper. The Davies-Bouldin and the Gap Statistic metrics are detailed in Subsections 2.2.1 and 2.2.2, respectively.

2.2.1 Davies-Bouldin

The characteristics of clustering that should be analyzed are the inter-clusters and intra-clusters dissimilarities. The problem is that one dissimilarity is penalized if the other is improved. In this way, Davies and Bouldin (Davies-Bouldin, db) proposed a metrics that balances both dissimilarities [16] [17]. This balance is illustrated in Equation 7. The combination of each two different clusters is analyzed internally and externally. The best number of groups should be the one that maximizes all two combinations using a range of possibility of numbers of clusters, exposed in Equation 6 where R is the maximum number of clusters. Each two combinations of clusters are compared to the intra-clusters distance divided by inter-clusters distance, as shown in Equation 7.

$$db = \frac{1}{R} \sum_{l=1, r=2, l \neq r}^{R} \max(db(l,r)), \tag{6}$$

$$db(l,r) = \frac{S_c(r) + S_c(l)}{d_{ce}(l,r)}. \tag{7}$$

The inter-clusters dissimilarity is the Euclidean distance between centers, represented in Equation 8. The intra-clusters dissimilarity is the Euclidean distance between the center $\mathbf{c}_r$ and the data grouped inside it, as shown in Equation 9 where N_r represents the total number of instances in the respective group.

$$d_{ce}(l,r) = ||\mathbf{c}_r - \mathbf{c}_l||. \tag{8}$$

$$S_c(r) = \frac{\sum_i ||\mathbf{x}_i - \mathbf{c}_r||}{N_r}, \tag{9}$$

2.2.2 Gap Statistics

Gap Statistics (Gap) was proposed by Tibishirani et al. [18] as an estimating metrics to select the number of clusters more efficiently than Hartigan [19], KL [20], CH [21] and Silhouette [22]. The advantage of the Gap is the evaluation of a random set in comparison to the dataset in the study. This set should have a behavior different from the expected.

Moreover, this metric measures the distance of all two combinations of data inside each cluster r as shown in Equation 10

$$D_r = \sum_{i=0, i'=1, i \neq i'}^{R} d_{ii'}, \tag{10}$$

where i and i' are the indexes of different instances inside the cluster, R is the maximum number of clusters, and $d_{ii'}$ is the Euclidean distance between the instances i and i'.

This combination provides a symmetric account of relation. For this reason, as can be seen in Equation 11, the result of Equation 10 should be divided by the double of all combinations between two instances, $2n_r$.

$$W_k = \sum_{r=1}^{R} \frac{1}{2n_r} D_r, \tag{11}$$

These calculations is done by a random uniform generated set (E_n^*) and by the current dataset. A random set shows an expected behavior of an input that is constrained by the problem limits. The Gap consequently is the subtraction of both behaviors represented in Equation 2.2.2, and this is the answer to be analyzed.

$$Gap_n(k) = E_n^*(log(W_k)) - log(W_k). \tag{12}$$

The choice of the number of groups can not be automatic with the answer because other aspects should be investigated: firstly, the standard deviation should be as lower as possible; secondly, the number of instances of the group contributes to understanding if the clustering process is forcing the division. Disproportional smaller groups indicate that they should not exist. An example is empty clusters. Thus, the choice of the number of groups is a complex problem not entirely comprehended in a metrics, but the Gap Statistics can indicate a promising solution.

Algorithm 1 provides the steps to be followed to the correct definition of the number of clusters according to the Gap Statistic. Observe that Equation is modified generating Equation 13, which allows the calculation of the metric.

```
Algorithm 1: Algorithm Gap
  1- Cluster the dataset X, varying the total number of clusters from $k = 1,2,...,K$, and determine $W_k$ using Equation 11;
  2- Generate $N_{ref}$ random uniformly distributed datasets, cluster them and calculate within dispersion measures $W^*_{k,ref}$,
     $ref = 1,2,...,N_{ref}$ to the values of $k = 1,2,...,K$, as done to the real data;
  3- Compute the estimated Gap Statistic for $k = 1,2,...,K$, according to Equation 13:
       $$Gap_k = \frac{1}{N_{ref}} \sum_{ref=1}^{N_{ref}} (log(W^*_{k,ref})) - log(W_k) \tag{13}$$
  4 - Let
       $$\bar{l} = \frac{1}{N_{ref}} \sum_{ref=1}^{N_{ref}} (log(W^*_{k,ref})) \tag{14}$$
  5- Compute the standard deviation using Equation 16:
       $$SD_k = \frac{1}{N_{ref}} \sum_{ref=1}^{N_{ref}} (log(W^*_{k,ref})) - \bar{l})^2]^{\frac{1}{2}} \tag{15}$$
  6 - Define
       $$S_k = SD_k\sqrt{1+\frac{1}{N_{ref}}} \tag{16}$$
  7 - Choose the number of clusters via:
       $$\hat{k} = \text{ smallest } k \text{ such that } Gap(k) \geq Gap(k+1) - S_{k+1} \tag{17}$$
```

3 Methodology

The main idea of this paper is to propose a methodology for clustering the students in an LMS to obtain better results than the original proposal presented in [3]. Thus, we used in this paper the same dataset of students used in [3]. The data utilized in the experiments is presented in [3]. This data belongs to the distance learning undergraduate course of Pedagogy of University of Pernambuco, Brazil [3]. 250 students compose this dataset were on the fourth semester, and 260 were in the fifth semester. Thus, we used two datasets, one dataset for each term (fourth semester and fifth semester). This Pedagogy course was selected because it is the one with the largest number of students and with the lowest dropout rate according to [3].

Both datasets (fourth semester and fifth semester) have 20 types of Portuguese grammar errors, which are described on Table 1. This table was redesigned based on [23]. This dataset of errors is a collection of common mistakes made during the process of writing. In both datasets, there are kinds of errors that were not committed by any students. For this reason, those errors were deleted from the datasets before the clustering process. Afterward, a normalization procedure was applied to avoid the privilege of any attribute in the clustering process. One can observe that each type of error is labeled with a particular number. Besides, some expressions in Portuguese can not be accurately translated to English, so they were left in Portuguese.

Once the dataset was defined and explained, we can describe the methodology used in this work. Oppositely to the proposal described in [3] that used only the K-means algorithm, we decide to use, besides K-means, another clustering algorithm called Fuzzy C-means (FCM). Thus, in our methodology, we used two well-known clustering algorithms, namely, K-means and FCM.

Table 1: The table describes all types of Portuguese grammar errors evaluated in the architecture proposed by [3].

Error Number	Category	Description
1	adv	Use of adverbs
2	adv — con	Agreement of adjective and noun
3	aha	Use of há/a
4	ali	Others
5	cjc	Use of conjunctions
6	cmt	Agreement between verbal modes and tenses
7	con	Nominal agreement
8	cop	Pronoun Placement
9	cop — pro	Use of mim and ti
10	cov	Verbal agreement
11	cov—reg	Verb Fazer
12	cra	Use of crase
13	ger	Use of gerund
14	mal	Use of mal/mau
15	pro	Use of pronouns
16	ptn	Use of punctuations
17	reg	Verbal regency
18	ren	Nominal regency
19	sem	Severe pleonasm
20	ver	Use of verbs

As it is well-known in the literature, the K-means has important drawbacks because the clusters formed depends on the initialization of centroids. Also, this algorithm is very sensible to noise and outliers. Because of this, we also decided to use the Fuzzy C-means algorithm to overcome the disadvantages of K-means algorithm in the problem tackled in this paper. Also, K-means and FCM algorithms are simple and adaptable to integrate with a real application.

Both K-means and C-means require a determination of the number K of clusters to be produced in the clustering process in advance. Therefore, the optimal number of clusters for the problem tackled in this paper needs to be determined. Accordingly, we evaluate a range of K (number of clusters) from two up to nine. We choose the best one according to a clustering performance metrics. Two metrics are pointed in the literature as relevant to be analyzed, Gap Statistics and Davies-Bouldin. We adopted both metrics in the analysis of this paper in addition to the standard dissimilarities intra and inter-clusters.

After the analysis of the best number of groups to be used in the clustering process, we execute the K-means and FCM using this number of clusters. Afterward, we can analyze the features of the groups of students formed considering their grammatical errors. A way to do this it is to evaluate the correlation between the characteristics (attributes) of each cluster of students. These correlations are analyzed with Spearman's correlation, which was chosen because it can be used in datasets with different distributions [4]. The existence of a high correlation between attributes within the clusters can help the professors to provide grammatical lessons for the students based on the association of the type of grammatical error. Thus, each student can be individually supported, what would favor the absorption of the content, and hence, improve the performance of the students.

4 Experiments and Results

In this section, we describe the results of the clustering process based on the grammar errors of the students. Table 2 presents the acronyms used to represent the datasets to be clustered. These id's are used to simplify the identification of the datasets used throughout the experiments. Each acronym is the junction of students' term (4th or 5th) and how many attributes are used (5, 6, 7, 8, 9, 11, 12) in each dataset. The experiments and results obeyed a chronologically ordered process, and the number of each subsection follows it. In Subsection 4.1, the first test analyzes both algorithms (K-means and FCM) and defines which one has a better performance. In Subsection 4.2, it is compared the results selecting the dimensions suggested by [3] and selecting all the dimensions which at least one student committed one mistake. In Subsection 4.3, the expressive attributes with high and different occurrences in the dataset are analyzed to check whether they should be maintained as inputs for clustering or not. In Subsection 4.4, the groups' dataset that preserved all prerequisites is clustered again to analyze the appearance of new correlations and similarities. In Subsection 4.5, the students are analyzed as groups. In the Subsection 4.6, we provide a final analysis of the whole process. Thus, the fuzziness coefficient utilized was 1.006, and this value was selected after previous experiments in which was observed that this value produced better results for the specific dataset.

Table 2: Acronyms of the datasets used in all the experiments.

Name	Description
4T12D	All the students in fourth term using 12 features.
4T11D	All the students in fourth term using 11 features.
4T7D	All the students in fourth term using 7 features.
4T6D	All the students in fourth term using 6 features.
5T9D	All the students in fifth term using 9 features.
5T8D	All the students in fifth term using 8 features.
5T6D	All the students in fifth term using 6 features.
5T5D	All the students in fifth term using 5 features.

4.1 Comparing the performance of K-means and Fuzzy C-means

The goal of this subsection is to analyze the capability of K-means and FCM to separate the students in clusters with high similarity among their participants and low similarity between members of different clusters. The analysis is based on the four metrics: Gap Statistics, Davies-Bouldin, Intra-cluster dissimilarity and Inter-cluster dissimilarity. In this paper, all the dissimilarity are calculated using Euclidean distance. The dissimilarity intra-cluster considers the mean of the sum of distances between each data and its centroid. Also, the dissimilarity inter-cluster calculates the average of the distances between each pair of centroids. Thus, we display the results of Davies-Bouldin metric using the logarithmic in order to facilitate the visualization of its behavior.

The dataset 4T12D is composed of all the students in the fourth term, and it is characterized by 12 attributes (**4T12D**), and the attributes with no counting of grammar errors for all the students were eliminated. The experimental configuration used in the first scenario was 30 simulations and 100 iterations, and the number of iterations was chosen based on the analysis of the convergence of the algorithms.

Figure 1 illustrates the results of each metric using K-means. Firstly, in Figure 1(a), the result of Gap Statistic does not show a good choice of numbers of clusters because the value continues to increase, but it does not demonstrate a high standard deviation. Secondly, in Figure 1(b), the results of Davies-Bouldin metric revealed the presence of uncertainty, once a high standard deviation is detected. So, the best choice of the number of clusters is visualized with 2 clusters. Thirdly, in Figure 1(c), the dissimilarity between centers exhibits the smallest standard deviation and the greatest value with 6 clusters. Lastly, in Figure 1(d), the dissimilarity inside clusters is minimized by 2 clusters, but it reveals a high standard deviation. All the metrics expose a high value and presence of standard deviation which defines the intrinsic uncertainty of K-means. Also, in several simulations, a number of group higher than 2 tends to create empty clusters. Based on these analyses, it is possible to assure that the best number of clusters for the 4T12D dataset is 2.

Figure 2 presents the results of the four metrics for the FCM algorithm. Figure 2(a) illustrates the behavior of the first metric: the Gap Statistic. The choice of 2 clusters is the best in this experiment, and it has the smallest standard deviation in comparison to the other possibilities for the number of clusters. Figure 2(b) shows the behavior of the Davies-Bouldin metric that presents high standard deviation in several values of numbers of clusters.

Figure 2(c) presents the Inter-clusters dissimilarity that exhibits a maximum value when the dataset is separated in 2 groups. Figure 2(d) shows the Intra-clusters dissimilarity that continues to minimize the distance using a large number of groups. Using FCM, the best choice is 2 clusters because it is the number that better appears in Gap Statistic and maximizes the difference between the centers.

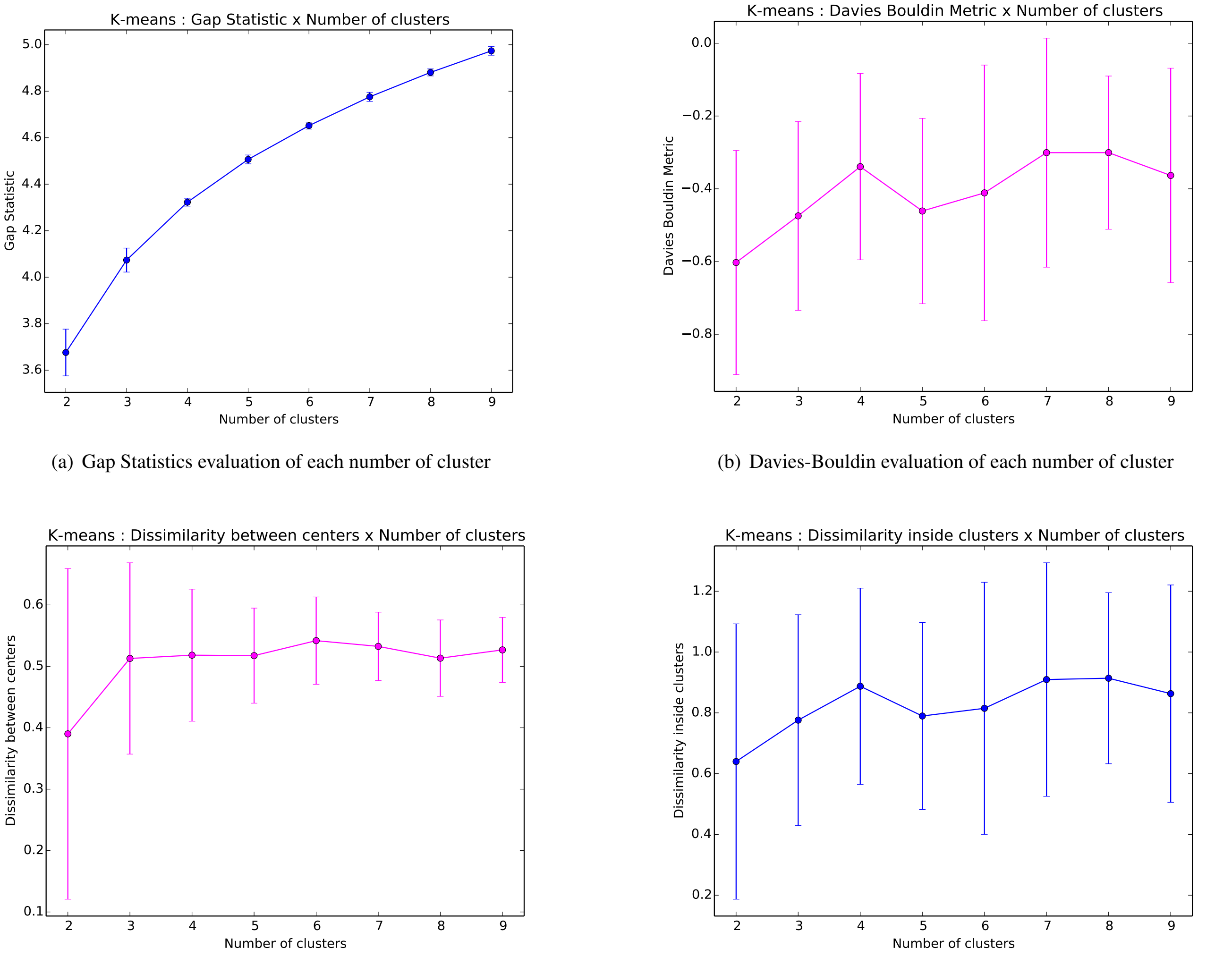
Fuzzy C-means compared to K-means demonstrated smaller standard deviations or, in other words, less uncertainty (more consistency) in the results. Moreover, K-means presented empty groups in the final clustering solution. In contrast, FCM had the capacity to separate the 4T12D dataset with any number of clusters. Consequently, FCM was selected to be used in the following experiments.

4.2 Comparing different numbers of attributes

In this subsection, all the four metrics are analyzed again using different numbers of attributes in two different cases. The first case is the 4T7D dataset that is composed by the students from the fourth term with the attributes whose mean of errors are greater than 0.05 (**4T7D** dataset). The 5T6D dataset used the same strategy to define it, but with the students from the fifth term (**5T6D**).

4.2.1 Fourth term

The **4T7D** dataset from the fourth term presents seven attributes for each pattern (student) that were selected using the new strategy of selection. To prove that the strategy for feature selection is adequate, we compared the performance of this strategy



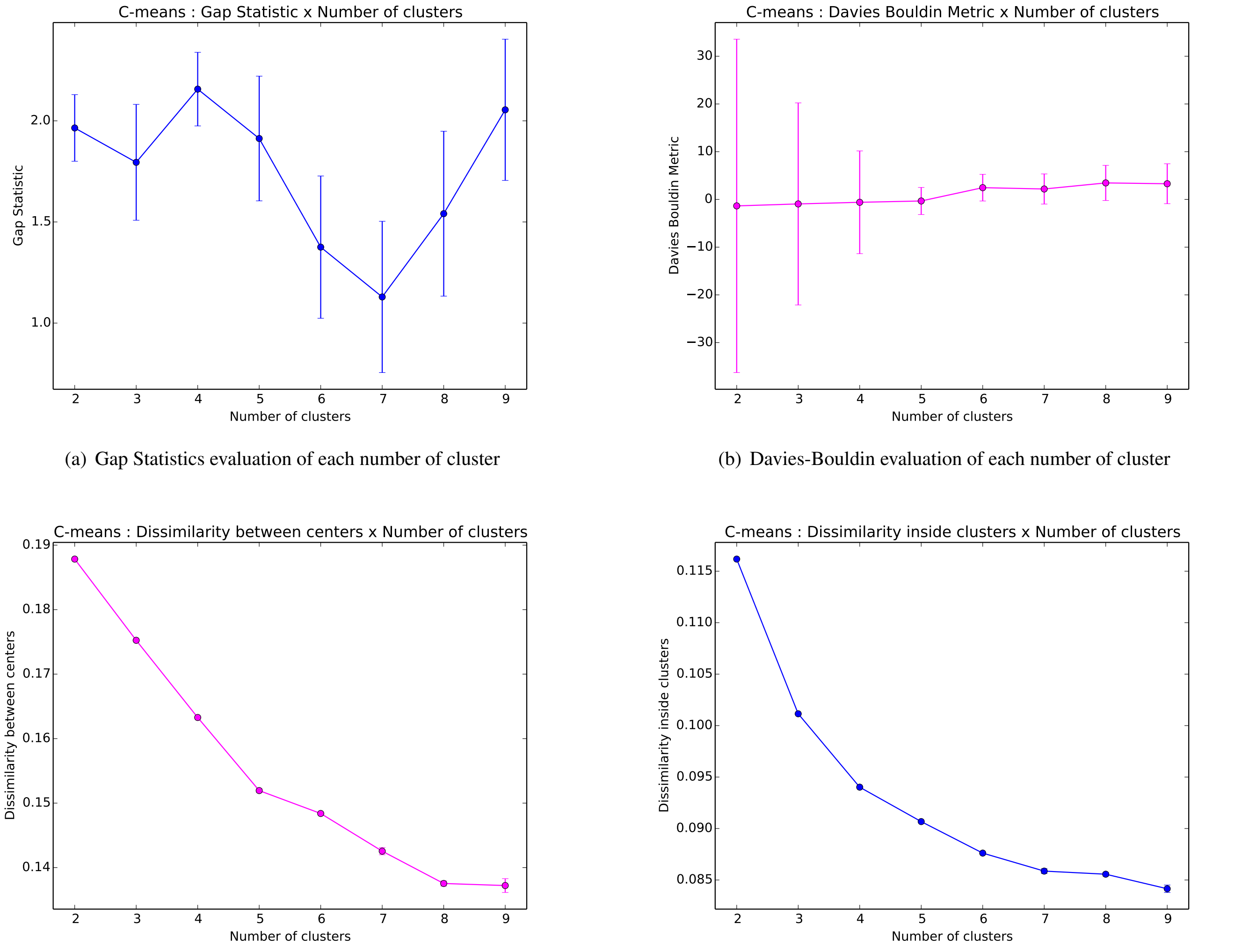
(a) Gap Statistics evaluation of each number of cluster

(b) Davies-Bouldin evaluation of each number of cluster

(c) Evaluation of dissimilarity between centers of each number of cluster

(d) Evaluation of dissimilarity inside groups of each number of cluster

Figure 1: Simulation results of **K-means** using 30 simulations and 100 iterations. The input was all the students from the fourth term with 12 attributes (types of errors), in other words, only the attributes with no counting of errors were deleted (**4T12D**).



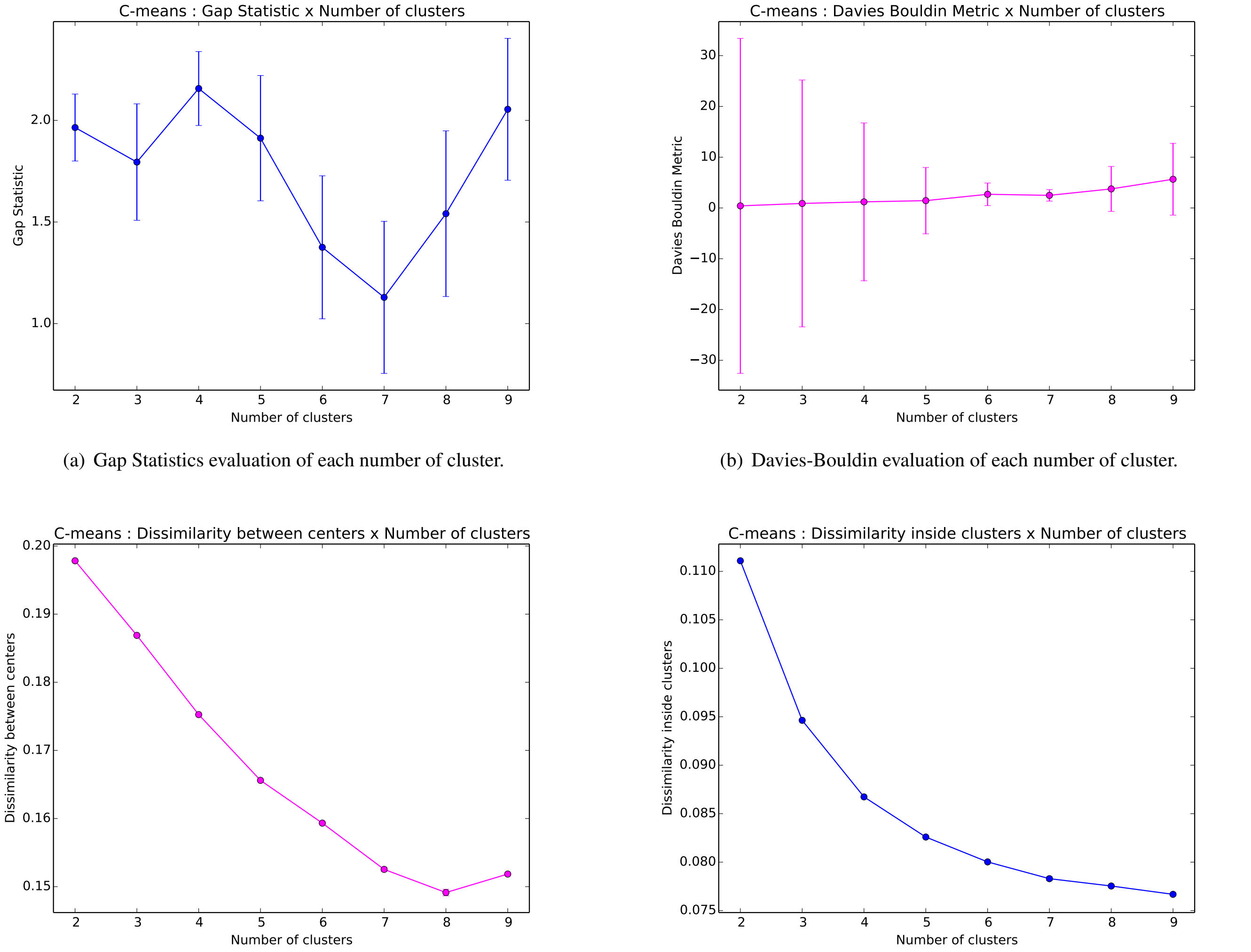
(a) Gap Statistics evaluation of each number of cluster

(b) Davies-Bouldin evaluation of each number of cluster

(c) Evaluation of dissimilarity between centroids of each number of cluster

(d) Evaluation of dissimilarity inside centroids of each number of cluster

Figure 2: Simulation results of **Fuzzy C-means** using 30 simulations and 100 iterations. The input was all the students from the fourth term with 12 attributes (types of errors), in other words, only the attributes with no counting of errors were deleted. (**4T12D**)



(a) Gap Statistics evaluation of each number of cluster.

(b) Davies-Bouldin evaluation of each number of cluster.

(c) Evaluation of dissimilarity between centroids of each number of cluster.

(d) Evaluation of dissimilarity inside centroids of each number of cluster.

Figure 3: Simulation results of C-means using 30 simulations and 100 iterations. The input was all the students of the fourth semester with seven dimensions (types of errors), in other words, the dimensions with means smaller than 0.05 and the empty ones for all the students were deleted **4T7D**.

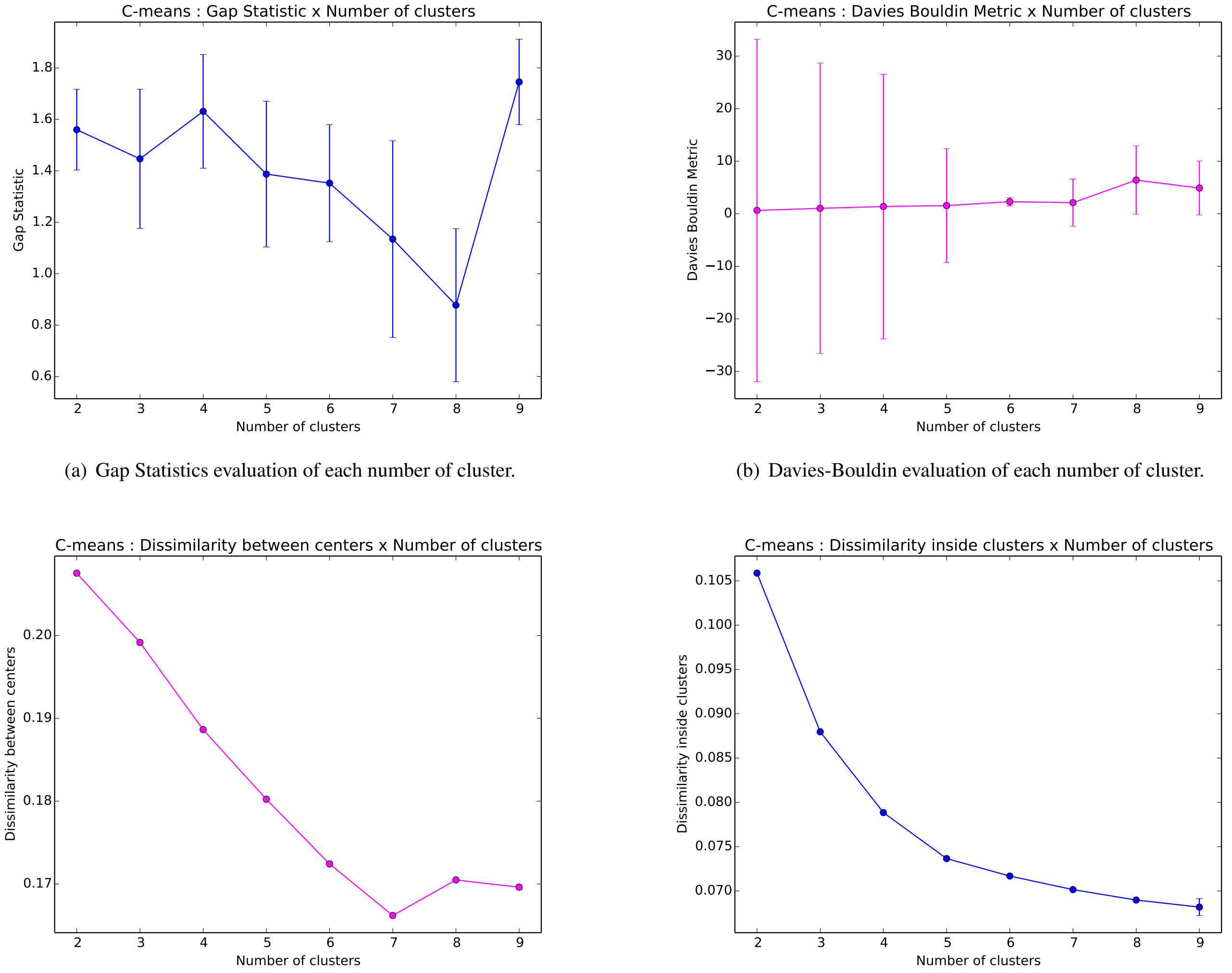
with the FCM considering the complete dataset without empty attributes (**4T12D**). The similarity for all metrics can be seen on the results of the dataset **4T12D**, in Figure 2, and **4T7D**, in Figure 3. The analysis of both results has the same explanation presented in Section 4.1. Gap Statistic reveals a better clustering with two groups. Davies-Bouldin expresses uncertainty using 2, 3 and 4 as numbers of clusters, and numbers greater than four present lower values of standard deviation. Both dissimilarities showed similar values in both experimental scenarios. Thus, this similarity proves that the attributes deleted by the feature selection strategy does not provide any relevant information for the clustering process, leading us to use the smaller dataset. The immediate gain is a lower computational cost (time) needed to perform the process.

4.2.2 Fifth term

The same methodology used to prove the difference between using or not the strategy of feature selection in the fourth term is utilized in the fifth term. In this way, the datasets **5T9D** and **5T6D** are compared to each other. It is possible to notice that both datasets presented similar results. The Gap Statistic is maximized with 2 clusters, whereas Davies-Bouldin is minimized with smaller numbers of clusters. Moreover, the intra-cluster dissimilarity values are equal or less than using more attributes, and the inter-cluster dissimilarity is greater or equal when compared. Furthermore, the eliminated attributes do not harm the results, and it reduces the computational cost. Consequently, the dataset **5T6D** is the best option to continue analyzing.

4.3 Comparing the relevance of a distinct attribute

Observing the database values, one particular attribute has values with a higher magnitude than others (Error number 7 - ”nominal agreement”, in Table 1). So, we perform an additional experiment using the dataset **4T7D** and **5T6D** to compare the



(a) Gap Statistics evaluation of each number of cluster.

(b) Davies-Bouldin evaluation of each number of cluster.

(c) Evaluation of dissimilarity between centroids of each number of cluster.

(d) Evaluation of dissimilarity inside centroids of each number of cluster.

Figure 4: Simulation results of C-means using 30 simulations and 100 iterations. The input was all the students of the fourth term with 6 dimensions (types of errors), in other words, the dimensions with means smaller than 0.05, dimensions with discrepant values in comparison to other dimensions and the empty ones were deleted **4T6D**.

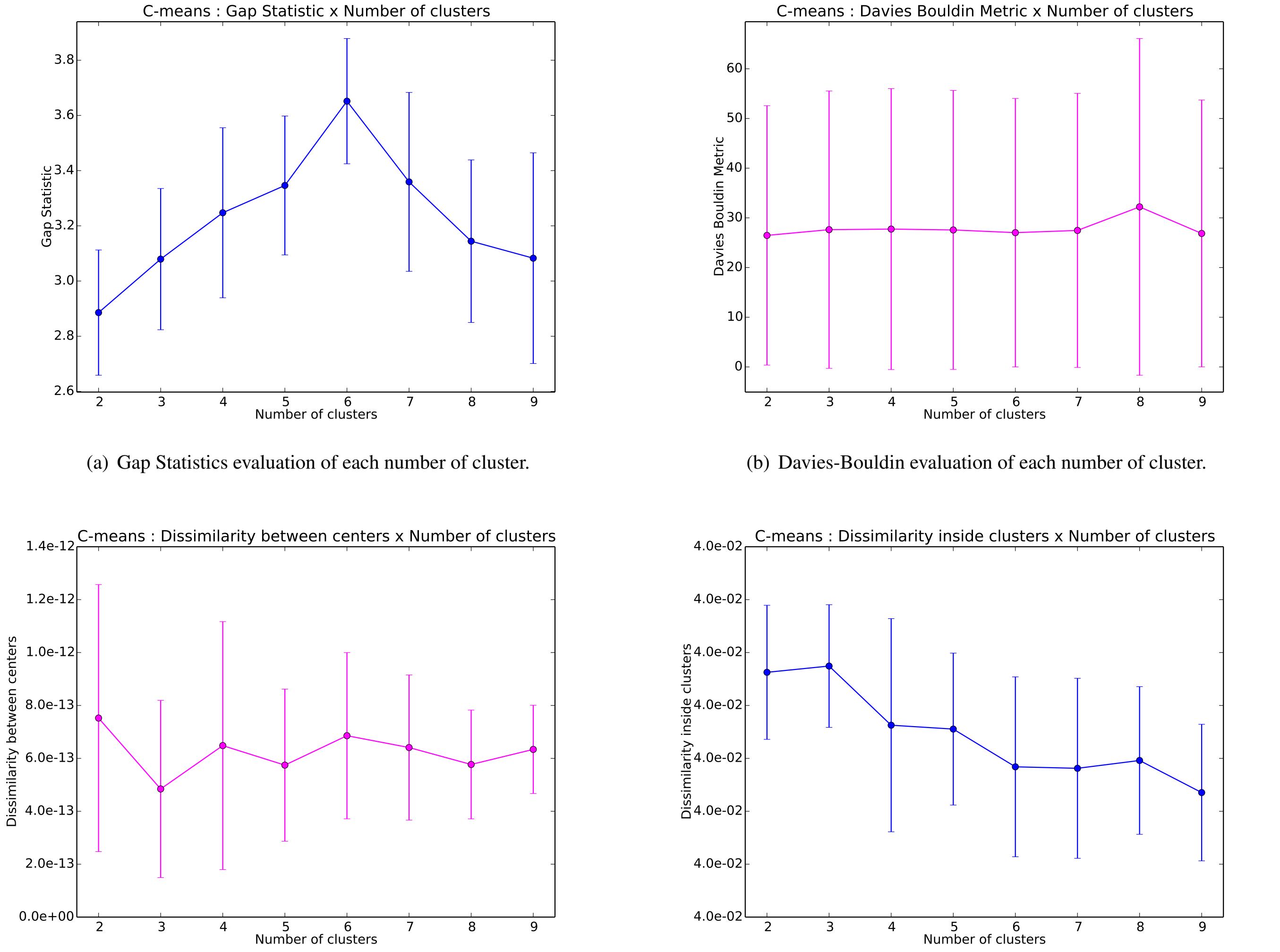
results with and without this particular attribute. This procedure yields the following datasets respectively: **4T6D** and **5T5D**. The hypothesis was if this distinct attribute avoids some similarities to emerge. In Figure 4, it is possible to see that all the metrics have higher presence of standard deviation than in Figure 3. Consequently, the removal of this distinct attribute harms the clustering process, increasing the uncertainty. In conclusion, the maintenance of this attribute leads to better results in the fourth term.

The same comparison was carried out in the fifth term. The standard deviation is increased in the **5T5D** compared to the **5T6D**. The values of inter-clusters dissimilarities raise, which means that each centroid is closer to the others. As mentioned, one of the primary goals is to have groups distant to each other and the removal of this attribute does not promote this behavior. In the fourth term, the harm is higher than in the fifth term. Therefore, the decision was to continue using this attribute inside the dataset.

4.4 Analyzing reclustered groups

Each cluster found by Fuzzy C-means demonstrated a low correlation between errors. It means that the students allocated in each group are not so similar as expected. This little relationship can be explained by the high diversity between the instances (patterns). However, the correlation can be increased with another independent clustering process for each group. Therefore, this experiment used the groups formed by both terms with better results (**4T7D** and **5T6D**).

In the fourth term, the new clustering of the first group obtained the results shown in Figure 5. As can be seen in Figure 5(a), the best number of clusters and the smallest standard deviation are found in 2 groups. Davies-Bouldin does not demonstrate a consistent result, once the standard deviation has large values. In addition, for the dissimilarity inter-cluster (Figure 5(c)), the best answer is with 2 groups, and for the intra-cluster (Figure 5(d)), a higher number of groups minimizes the distance internally.



(a) Gap Statistics evaluation of each number of cluster.

(b) Davies-Bouldin evaluation of each number of cluster.

(c) Evaluation of dissimilarity between centroids of each number of cluster.

(d) Evaluation of dissimilarity inside centroids of each number of cluster.

Figure 5: Simulation results of C-means using 30 simulations and 100 iterations. The input was the cluster 1 of all the students in the fourth semester with 6 dimensions (types of errors), in other words, the dimensions with means smaller than 0.05, dimensions with discrepant values in comparison to other dimensions and the empty ones were deleted.

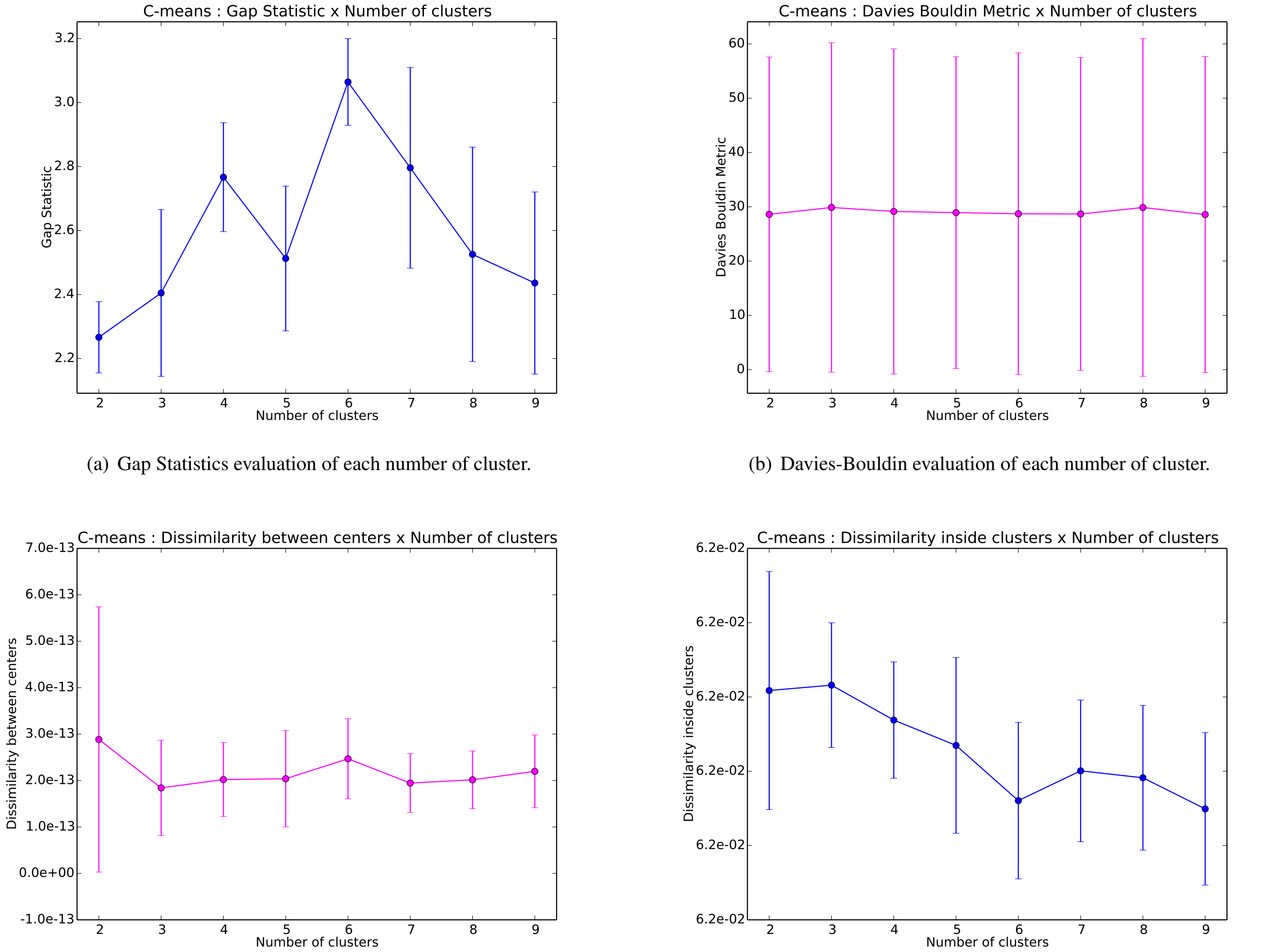
The majority of the metrics indicates that 2 groups is the best choice.

The second group of the fourth term resulted in the outcomes showed in Figure 6. In Figure 6(a), Gap Statistic has the best result with two groups combined with one of the smallest standard deviation values, observing that two fits in a better way because it is the least number of groups which provide a satisfactory answer. In Figure 6(b), the best answer of Davies-Bouldin produced a high standard deviation for all number of groups. In Figure 6(c), the number of 2 groups continues to be the best option, and, in Figure 6(d), a large number of groups presents better responses. This behavior can be explained by the fact that the highest number of groups provides empty clusters and, consequently, it decreases the intra-cluster distance. As the goal is to separate the data, the result in Figure 6(d) is not relevant in comparison to the others, and the best option is 2 clusters. Using the clustered groups of the fifth term, the metrics demonstrate that the clustering with two groups is the best option for both sub-clusters.

4.5 Analyzes of Grammatical Errors of the Students

The goal in this section is to analyze the students inside each group, their characteristics, and correlations. After several experiments, the clustering that should be analyzed are the ones of the fourth and fifth terms using 7 and 6 attributes, respectively. All the experimental scenarios showed that it was necessary to divide the students into two groups. Comparing the results, it is possible to identify that several students moved to other clusters. The groups are not similar to each other.

In the first clustering of **4T7D**, it can be observed in Table 3 and 4 that the majority of the correlations between the attributes inside each cluster is small. Same correlations invert or intensify their values, meaning a difference between the groups. The correlation of Error 7 and 10 seems to be present in both groups, revealing the necessity of creating recommendations of lessons associated with these topics for all the students. Errors 10 and 16 and Errors 7 and 8 indicate a distinct characteristic between



(a) Gap Statistics evaluation of each number of cluster.

(b) Davies-Bouldin evaluation of each number of cluster.

(c) Evaluation of dissimilarity between centroids of each number of cluster.

(d) Evaluation of dissimilarity inside centroids of each number of cluster.

Figure 6: Simulation results of C-means using 30 simulations and 100 iterations. The input was the cluster 0 of all the students in the fourth term with 6 dimensions (types of errors), in other words, the dimensions with means smaller than 0.05, dimensions with discrepant values in comparison to other dimensions and the empty ones were deleted.

groups. Students need different recommendations of lessons on these topics.

In the second clustering of **4T7D**, in Tables 5-8, same correlations are close to 0.0, meaning the non-existence of correlation between these attributes. An example of this is the Errors 12 and 13, where the correlation is around 0.0001 for one cluster and -0.14 for the other. This difference helps to avoid recommendations of lessons relating these two grammatical errors.

In the first clustering of **5T6D**, the correlation within the clusters can be compared to each other. The relation between Error 7 and 10 continues to be higher than the others for both groups. These errors illustrate a lack of correlation for one of the clusters. In the second clustering of **5T6D**, same correlations are intensified. Errors 10 and 13 or 10 and 12 for one cluster are intensified, meaning a correlation of these errors for some students.

4.6 Final analysis

In [4], Davies-Bouldin had good results compared to Gap Statistic. However, in the real dataset used in that paper, the mass of data is harder to be separated. It should be analyzed the dissimilarity intra and inter-clusters instead.

In [3], the number of clusters for a student grammatical errors dataset after the clustering was 3 and the algorithm used in clustering was K-means. In this paper, it is proved that the Fuzzy C-means (FCM) should be used instead because it yields better clustering solutions and that this dataset has the tendency to be grouped into 2 clusters according to the methodology employed in this paper. Finally, this paper demonstrates how clustering techniques can be applied as a tool for suggesting lessons for students with grammatical difficulties according to the association of correlated errors. Hence, students can be supported with appropriate material with a combination of relevant topics to avoid studying unnecessary subjects.

It is worth mentioning that this paper provides a computational and statistical methodology for each decision making process accomplished in the clustering process of the student grammatical errors dataset. The first step of our methodology showed that

	Error 1	Error 7	Error 8	Error 10	Error 12	Error 13	Error 16
Error 1	1.0	0.1583	-0.0473	0.1720	0.08883	0.2128	-0.0604
Error 7	0.1583	1.0	0.0942	**0.2972**	0.2422	0.1853	0.0448
Error 8	-0.0473	0.0942	1.0	-0.0332	-0.0101	-0.0988	-0.0446
Error 10	0.1720	**0.2972**	-0.0332	1.0	0.0725	0.1099	0.04093
Error 12	0.0888	0.2422	-0.0101	0.0725	1.0	0.1271	0.0395
Error 13	0.2128	0.1853	-0.0988	0.1099	0.1271	1.0	0.0367
Error 16	-0.0604	0.0448	-0.0446	0.0409	0.0395	0.03677	1.0

Table 3: This table detailed the correlations inside the first group of the first clustering using **4T7D** dataset.

	Error 1	Error 7	Error 8	Error 10	Error 12	Error 13	Error 16
Error 1	1.0	-0.0503	0.0478	0.0161	0.0399	0.0842	-0.1508
Error 7	-0.0503	1.0	0.0688	0.1630	0.3336	0.0539	0.0247
Error 8	0.0478	0.0688	1.0	-0.0244	0.0538	0.0869	-0.1243
Error 10	0.0161	0.1630	-0.0244	1.0	0.2867	**0.4158**	**0.3313**
Error 12	0.0399	**0.3336**	0.0538	0.2867	1.0	**0.0001**	0.1889
Error 13	0.0842	0.0539	0.0869	**0.4158**	**0.0001**	1.0	0.1412
Error 16	-0.1508	0.0247	-0.1243	0.3313	0.188	0.1412	1.0

Table 5: This table detailed the correlations inside the first group of the second clustering using **4T7D** dataset and group one.

	Error 1	Error 7	Error 8	Error 10	Error 12	Error 13	Error 16
Error 1	1.0	0.0679	-0.0553	0.1972	0.1183	0.1480	-0.0126
Error 7	0.0679	1.0	0.0966	0.0753	0.1247	0.0746	0.0361
Error 8	-0.0553	0.0966	1.0	-0.1397	-0.0414	-0.0930	0.0362
Error 10	0.1972	0.0753	-0.1397	1.0	0.1478	-0.0248	-0.0185
Error 12	0.1183	0.1247 2	-0.0414	0.1478	1.0	0.1444	0.1043
Error 13	0.1480	0.0746 4	-0.0930	-0.0248	0.1444	1.0	0.1166
Error 16	-0.0126	0.0361	0.0362	-0.0185	0.1043	0.1166	1.0

Table 7: This table detailed the correlations inside the first group of the second clustering using **4T7D** dataset and group two.

	Error 1	Error 7	Error 8	Error 10	Error 12	Error 13	Error 16
Error 1	1.0	0.1249	0.0429	-0.0384	-0.0202	0.1603	0.0678
Error 7	0.1249	1.0	0.2225	**0.3999**	0.1299	0.1037	0.1776
Error 8	0.0429	0.2225	1.0	0.0301	0.1260	0.0939	0.0489
Error 10	-0.0384	**0.3999**	0.03019	1.0	0.0956	0.1799	0.2583
Error 12	-0.0202	0.1299	0.1260	0.0956	1.0	-0.0709	0.0155
Error 13	0.1603	0.1037	0.0939	0.1799	-0.0709	1.0	0.1295
Error 16	0.0678	0.1776	0.0489	0.2583	0.0155	0.1295	1.0

Table 4: This table detailed the correlations inside the second group of the first clustering using **4T7D** dataset.

	Error 1	Error 7	Error 8	Error 10	Error 12	Error 13	Error 16
Error 1	1.0	0.0617	-0.0305	-0.2071	-0.0878	0.1970	0.2069
Error 7	0.0617	1.0	**0.0040**	0.4084	0.0158	0.1071	0.1163
Error 8	-0.0305	**0.0040**	1.0	-0.0822	0.1308	0.0828	0.1526
Error 10	-0.2071	0.4084	-0.0822	1.0	-0.1097	0.0177	0.1690
Error 12	0.0878	0.0158	0.1308	-0.1097	1.0	-0.1426	-0.1193
Error 13	0.1970	0.1071	0.0828	0.0177	-0.1426	1.0	0.0900
Error 16	0.2069	0.1163	0.1526	0.1690	-0.1193	0.0900	1.0

Table 6: This table detailed the correlations inside the second group of the second clustering using **4T7D** dataset and group one.

	Error 1	Error 7	Error 8	Error 10	Error 12	Error 13	Error 16
Error 1	1.0	0.1296	-0.0618	0.0492	0.0392	0.2217	-0.1157
Error 7	0.1296	1.0	0.0102	-0.1601	0.1737	0.08282	**-0.2544**
Error 8	-0.0618	0.0102	1.0	-0.0572	-0.0143	-0.1293	-0.1088
Error 10	0.0492	-0.1601	-0.0572	1.0	-0.1178	0.0786	-0.0328
Error 12	0.0392	0.1737	-0.0143	-0.1178 5	1.0	0.0725	-0.0325
Error 13	0.2217	0.0828	-0.1293	0.0786	0.0725	1.0	-0.0333
Error 16	-0.1157	**-0.2544**	-0.1088	-0.0328	-0.0325	-0.0333	1.0

Table 8: This table detailed the correlations inside the second group of the second clustering using **4T7D** dataset and group two.

the performance of K-means had a behavior with the presence of a high standard deviation, which means that it does not provide consistent results. In contrast, FCM provides smaller values and standard deviation.

In the second step, two types of selected attributes were compared to each other. In this case, the removed attributes were not relevant to the clustering process. For this reason, we chose the economic option.

In the third step, the presence of an attribute composed with higher values in comparison to the others was tested as a problematic attribute to the clustering of the dataset. The removal of this attribute brought uncertainty to the results attesting that this action was not beneficial. Next, using all the decisions in previous experimental scenarios, the clustering is applied in both clusters of students formed in initial steps of our methodology. The second clustering brought the intensification or weakening of some correlations between attributes. Moreover, the average of intra-cluster distance in each group is decreased. This advance causes more specification of the different behaviors and types of students. Finally, the students can be assisted with special materials generated by specific characteristics of the overall difficulties and specific ones.

At last, the results were analyzed by an educational specialist who agreed that in real practice it is possible to observe the high occurrence of the errors 7 and 10 simultaneously, and, therefore, there is a strong correlation between them. The specialist also noticed that the re-clustering of one group presented an intensification of the correlation between the errors 7 and 12 which is also observed in real practice. Thus, these correlations and others can provide the elaboration of specific recommendations of grammatical lessons regarding the particular difficulties of each student.

The re-clustering of a dataset is not the same of a clustering using the higher number of groups, for example, the clustering of a dataset in 4 groups is not the same of clustering a dataset into two groups and re-clustering both groups again in two groups. This difference is because when some instances are taken away from a dataset, some relationships are highlighted, and others are weakened. Consequently, the division of a highly diverse data is different from a more selected group because it is based on various behaviors. The re-clustering may be added as a step because the result will affirm when the data cannot be divided anymore. In each re-clustering, the result turns into a narrow and specific analysis.

5 Conclusions

This paper investigated the application of two widely used clustering algorithms - K-means and Fuzzy C-Means (FCM) - to group students based on their grammatical errors collected on an on-line educational platform of a Brazilian university. The goal was elaborates a tool to increase the performance of a recommendation system in view of improve the learning process of the students.

K-means is the most used clustering algorithm in the literature because of its simplicity and low computational cost requirement. However, it is very susceptible to initialization and cannot separate overlapping data. FCM can overcome these difficulties by means of the elaboration of a pertinence matrix. Besides, it was tested different numbers of final groups and four validation metrics to compare the performances: GAP statistics, Davies-Bouldin, intra-cluster and inter-cluster distances.

In the first stage of this investigation we use both methods to cluster the students of the fourth term of the Pedagogy undergraduate course, considering various numbers of attributes of the dataset. Fuzzy C-means could separate the students maintaining small standard deviation in dissimilarities intra and inter-cluster. Based on these results we perform the clustering of the students of the fifth term to define the best number of attribute. Then, we re-cluster the formed groups in view of to increase the quality of the clusters.

Spearman's correlation showed the absence of relationship among errors made by students, revealing the necessity of specific recommendations for study materials to improve each one of these grammatical difficulties. Same errors were not strongly related. However, it could be done a combination of suggestions of exercises to create diversity in the materials and to avoid boredom in the students while doing them.

Besides the high amount of metrics in the literature, they continue to be only a clue of which a number of groups are better to cluster an uncategorized dataset. The metrics do not penalize empty clusters and inconstancy. Also, Davies-Bouldin demonstrated that it is not a good metric to be used in a diversified dataset.

This paper reveals the complexity of choosing strategies to continue investigating students' similarities and necessities. Moreover, it intensifies the importance of the methodology used in this paper. As future works, other algorithms based on swarm intelligence can be applied to solve these and others clustering tasks related to Educational data.

References

[1] C. Romero and S. Ventura. "Data mining in education". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12–27, 2013.

[2] A. Peña-Ayala. "Educational data mining: A survey and a data mining-based analysis of recent works". *Expert systems with applications*, vol. 41, no. 4, pp. 1432–1462, 2014.

[3] F. B. Soares. "Desenvolvimento de uma arquitetura utilizando Mineração de dados educacionais para apoiar alunos de EAD com dificuldades na gramática portuguesa ". Master's thesis, University of Pernambuco, 2017.

[4] M. G. Macedo and C. J. Bastos-Filho. "Clustering Users Based on the Capacity to Solve Questions in an Educational Platform". *XIII Encontro Nacional de Inteligência Artificial e Computacional*, pp. 121–132, 2016.

[5] E. W. Forgy. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". *Biometrics*, vol. 21, pp. 768–769, 1965.

[6] S. Lloyd. "Least squares quantization in PCM". *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[7] J. MacQueen *et al.*. "Some methods for classification and analysis of multivariate observations". In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 281–297. Oakland, CA, USA., 1967.

[8] J. A. Hartigan and M. A. Wong. "Algorithm AS 136: A k-means clustering algorithm". *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[9] X. Lan, Q. Li and Y. Zheng. "Density K-means: A new algorithm for centers initialization for K-means". In *Software Engineering and Service Science (ICSESS), 2015 6th IEEE International Conference on*, pp. 958–961. IEEE, 2015.

[10] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.

[11] S. Ghosh and S. K. Dubey. "Comparative analysis of k-means and fuzzy c-means algorithms". *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 4, 2013.

[12] J. C. Bezdek, R. Ehrlich and W. Full. "FCM: The fuzzy c-means clustering algorithm". *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[13] Y. Yong, Z. Chongxun and L. Pan. "A novel fuzzy c-means clustering algorithm for image thresholding". *Measurement Science Review*, vol. 4, no. 1, pp. 11–19, 2004.

[14] S. Chattopadhyay, D. K. Pratihar and S. C. De Sarkar. "A comparative study of fuzzy c-means algorithm and entropy-based fuzzy clustering algorithms". *Computing and Informatics*, vol. 30, no. 4, pp. 701–720, 2011.

[15] A. Stetco, X. jun Zeng and J. Keane. "Fuzzy C-means++: Fuzzy C-means with effective seeding initialization". *Expert Systems with Applications*, vol. 42, no. 21, pp. 7541–7548, 11 2015.

[16] D. L. Davies and D. W. Bouldin. "A cluster separation measure". *IEEE transactions on pattern analysis and machine intelligence*, , no. 2, pp. 224–227, 1979.

[17] J. Vesanto and E. Alhoniemi. "Clustering of the self-organizing map". *IEEE Transactions on neural networks*, vol. 11, no. 3, pp. 586–600, 2000.

[18] R. Tibshirani, G. Walther and T. Hastie. "Estimating the number of clusters in a data set via the gap statistic". *J. R. Statist. Soc. B*, vol. 63, no. 2, 2001.

[19] J. Hartigan. *Clustering algorithms*. John Wiley and Sons, Inc, 1975.

[20] W. Krzanowski and Y. Lai. *A criterion for determining the number of groups in a data set using sum-of-squares clustering*. Biometrics, 1988.

[21] T. Calinski and J. Harabasz. *A dendrite method for cluster analysis*. Communications in Statistics-Theory and Methods, 1974.

[22] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, Canada, 1990.

[23] W. D. C. d. M. Silva. "Aprimorando o corretor gramatical CoGrOO. 2013. 166f". Ph.D. thesis, Dissertação (Mestrado em Ciência da Computação). Instituto de Matemática e Estatística, Universidade de São Paulo. São Paulo, 2013.