Applied Computing and Informatics 15 (2019) 153–162

Contents lists available at ScienceDirect

Applied Computing and Informatics

journal homepage: www.sciencedirect.com

Original Article

Research on social data by means of cluster analysis

Camila Maione [a], Donald R. Nelson [b], Rommel Melgaço Barbosa [a,*]

[a] Instituto de Informática, Universidade Federal de Goiás, Goiânia, Goiás, Brazil
[b] Department of Anthropology, University of Georgia, Athens, USA

ARTICLE INFO

Article history:
Received 13 December 2017
Revised 20 February 2018
Accepted 20 February 2018
Available online 21 February 2018

Keywords:
Clustering
Social data
Classification
Pam
Data mining

ABSTRACT

This paper presents a data mining study and cluster analysis of social data obtained on small producers and family farmers from six country cities in Ceará state, northeast Brazil. The analyzed data involve demographic, economic, agriculture and food insecurity information. The goal of the study is to establish profiles for the small producer families that reside in the region and to identify relevant features which differentiate these profiles. Moreover, we provide an efficient data mining methodology for analysis of social data sets which is capable of handling its natural challenges, such as mixed variables and abundance of null values. We use the Silhouette method for the estimation of the best number of natural groups within the data, along with the Partitioning Around Medoids clustering algorithm in order to compute the profiles. The Correlation-Based Feature Selection method is used to identify which social criteria are the most important to differentiate the families from each profile. Classification models based on support vector machines, multilayer perceptron and decision trees were developed aiming to predict in which of the identified clusters an arbitrary family would be best fit. We obtained a good separation of the families into two clusters, and a multilayer perceptron model with approximately 93.5% prediction accuracy.

1. Introduction

The Brazilian law (Lei 11.326, 2006) defines the family farmer as the family which practices agricultural activities in a rural area that is not larger than four fiscal modules and uses predominantly the family's own labor in their business activities. Moreover, according to the law, the income of a family farmer must be originated from activities related to their own business, which must be ran by the farmer and his family. According to Guanziroli [15], in 2001, family-based agriculture in Brazil was so far responsible for almost 40% of the total agricultural production, and accounts for 76.8% of agricultural employment.

Most of the family-based agriculture in Brazil is consolidated in the Northeast region. However, the family farmers which reside in this region face various challenges related to the soil and climate. Approximately 70% of the semiarid region is above the crystalline basement where soils are generally shallow and with low water infiltration capacity, which limit the development of agriculture [16]. Furthermore, the Northeast region frequently suffers from drought, resulting in insufficient or irregular distribution of rainfall, which cannot keep the soil moist enough during periods longer than the harvest cycle [8]. In addition, the income distribution among the farmer families in the region appears to be deficient. In 2006, family-based agriculture in the Northeast region was composed of more than 4,500,000 farmers; only around 450,000 of them comprised the most capitalized group, with a annual income of more than 53,000 reais, while more than 2,500,000 farmers survived only on subsistence activities, with an annual monetary income of approximately 255 reais [16].

One possible way to identify categories among these families and to gain insight on the most important factors for this division is the analysis of social data regarding these families by cluster analysis and data mining techniques. Data mining is a process of discovery of useful information and hidden patterns in databases, widely used in recent research in various fields such as food science [3,4,26,27], agriculture [28,29], social media [2,17], business and customer management [30], sports [6] and others. However, the analysis of social data sets present some challenges derived from the manner in which they are gathered and stored. Social data sets

* Corresponding author address: Alameda Palmeiras, Quadra D, Câmpus Samambaia, Instituto de Informática, CEP 74690-900 Goiânia, Goiás, Brazil.
E-mail address: rommel@inf.ufg.br (R.M. Barbosa).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.aci.2018.02.003
2210-8327/

are composed of many mixed variables, which restrain the direct application of one the most popular clustering algorithms, the K-means. There is also a large quantity of null values distributed among the data that must be handled. Due to the exploratory nature of the analysis that we wish to conduct, the number of natural groups existing within the data is unknown a priori, and many clustering techniques require this value as an input parameter.

This study presents a data mining study which aims to establish profiles for the family farmers residing in Northeastern Brazil. The proposed methodology include data mining, cluster analysis and variable selection techniques in order to identify profiles among the analyzed families based on demographic, economic, agricultural production and food insecurity information. We focus our study on family farmers from six municipalities from the state of Ceará. We identified homogeneous groups within the available data, aiming to find which families are relatively similar in terms of the gathered information and which of these social criteria are more relevant for the differentiation of the family groups. We built classification models capable of predicting in which of these groups an arbitrary family would be best fit. Decision rules were employed to aid the interpretation of the clustering results.

Finally, this paper presents the following contributions:

- We conducted a data analysis of social information in order to identify profiles of small farmer families from the state of Ceará, using statistical and machine learning techniques from data mining and cluster analysis;
- We provide a data mining methodology which is capable of handling challenges inherent to social data sets, such as mixed variables, abundant null values and the absence of information about the number of natural groups within the data.

The content of this paper is organized as follows: Section 2 describes the methodology employed, the data set, concepts and algorithms used; Section 3 details the results obtained; Section 4 brings the conclusion.

2. Methodology

The analyzed data was obtained from research conducted on 476 households distributed around six country cities from the Ceará state: Barbalha, Guaraciaba do Norte, Boa Viagem, Limoeiro do Norte, Itarema and Parambu (Fig. 1). Each family was evaluated by a questionnaire composed of 79 questions regarding demographic, economic, agricultural production and food security information. The answers for the questions were defined as the features for the data set. Some important questions (features) are listed in Table 1. The number of evaluated families from each city is similar: 80 from Barbalha, 80 from Boa Viagem, 80 from Guaraciaba do Norte, 79 from Itarema, 79 from Limoeiro do Norte and 78 from Parambu. Data were collected in 2012 over a 1-month period, as part of a research project focused on understanding household drought vulnerability. The six municipalities selected each represent one of the agro-ecological zones in the state. Questions were developed based on a sustainable livelihoods approach [34], which categorizes five types of capital – social, natural, financial, human and physical – and provides a framework for understanding their relationships. The respondents were randomly selected from a list of farmers that were provided by the local Farm Workers Syndicate, which contains an almost exhaustive list of farmers Table 2.

2.1. Knowledge discovery using data mining and cluster analysis

The first step of the analytical procedure was to identify relevant groups of the interviewed families based on a similarity factor related to the nature and domain of the social questions involved. For this matter, we employed cluster analysis concepts and techniques.

Data mining refers to the process of automatic discovery of useful information in large databases [35]. Data mining techniques merge machine learning, statistical calculation, linear algebra and mathematic optimization concepts in order to uncover hidden patterns in data. Machine learning is a fundamental procedure for data mining processes as it generates smart algorithms capable of discovering patterns and information in data bases automatically, which aid decision making [5].

Cluster analysis is a data mining process which consists in dividing the samples into groups (clusters) based on information found within the data which describes these samples and its relationships [35]. Samples belonging to the same cluster must show a similarity pattern among them while being as dissimilar as possible from samples associated to other clusters.

Due to the nature of social data and the way they are usually obtained and stored, the application of clustering techniques on this kind of data proves challenging. Research on social data is usually conducted through the use of structured forms which bring a number of questions to be answered by the members of the observed population. Once collected, these data are stored in spreadsheets or plain text. The conversion of these files to data matrixes usually generates columns of mixed types, including numeric, categoric and ordinal variables. Moreover, fields that are not filled are mapped into null values, which represent obstacles for data analysis.

Some recent research brings methodologies and techniques for handling the problem of data sets with variables of mixed types during cluster analysis [1,21,32]. Another point to be observed is that, in this study, we conduct an exploratory analysis where we have no pre-established models or hypotheses regarding the data [22]. Most clustering techniques require the number of clusters as an input parameter, and in our case this number is not known a priori. We employed the Silhouette method which aids in the estimation of the best number of clusters within the data set.

2.1.1. Estimating the best number of clusters

In this study, we used the Silhouette method [33] to estimate the best number of natural clusters within the data set. Silhouette is a graphical display method for clustering algorithms. Each cluster is represented by a silhouette which is designed based on similarity and dissimilarity between the samples. The silhouette shows which samples are well placed in the clusters and which samples are floating between two or more clusters. The entire partition is shown in a single plot as a combination of silhouettes which allows visualization of the quality of the clusters. The best number of clusters can be estimated by computing the average weigth of the silhouettes designed for a range of partitions with different numbers of clusters.

Consider $a(i)$ as the average distance between the sample i and the other samples in the same cluster as i. For all other clusters C_k, k being the number of clusters, $d(i, C)$ is the average distance between i and the other samples in C_k. The minimum value of $d(i, C)$, given by $b(i) := \min_c d(i, C)$, is the distance between i and the nearest cluster. Hence, $s(i)$ is computed as:

$$s(i) := \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Samples with show high values for $s(i)$ are well clustered, otherwise they are probably placed between two or more clusters, and samples with negative values for $s(i)$ are placed in wrong clusters. The quality of a cluster can be estimated by computing the average value of $s(i)$ for all samples associated to it.

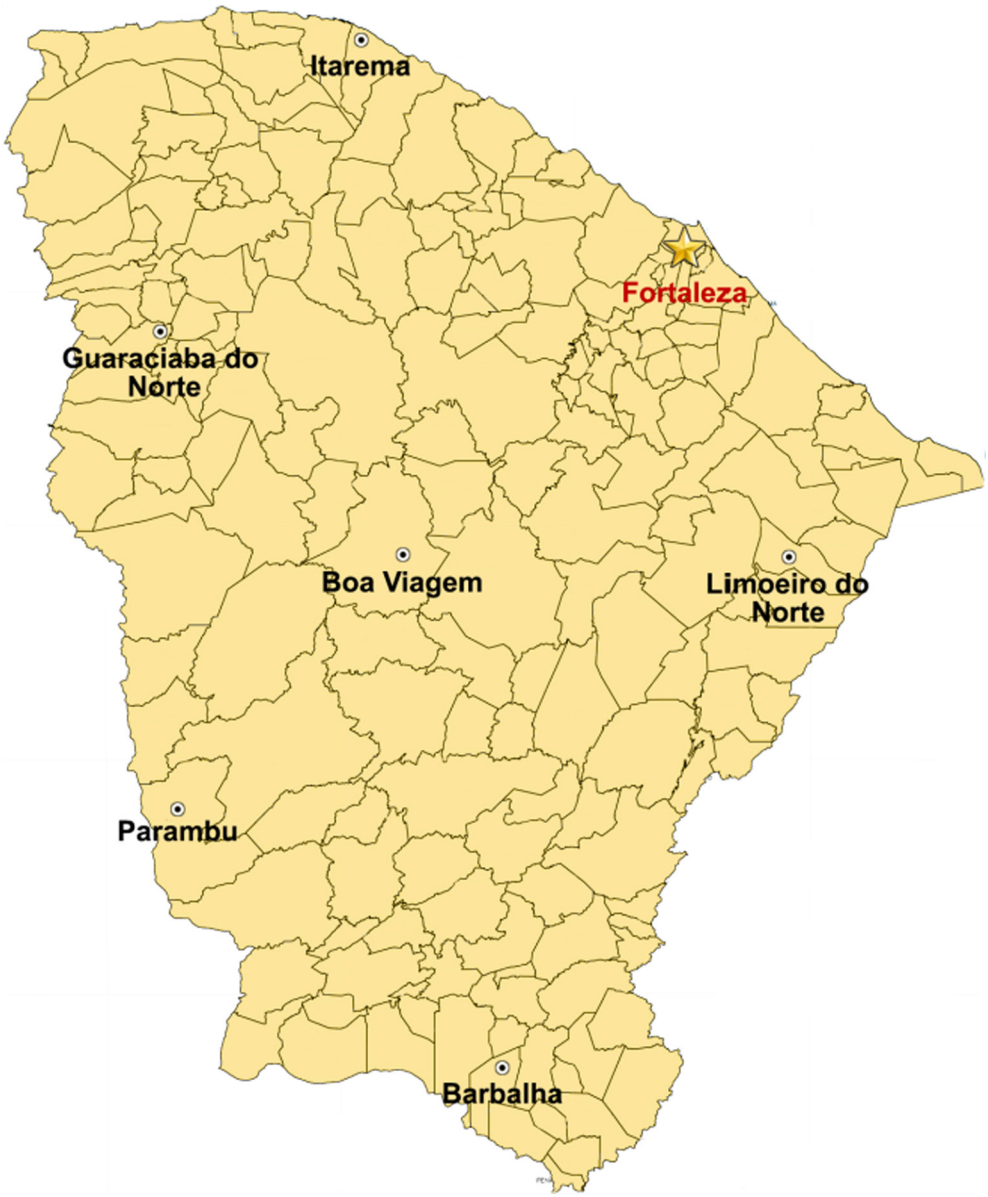

Fig. 1. Map of Ceará state. The six cities of origin of the family farmers analyzed (Barbalha, Guaraciaba do Norte, Boa Viagem, Limoeiro do Norte, Itarema and Parambu) and the capital, Fortaleza, are highlighted.

In order to choose the best number of clusters, we selected a range of possible values, $F = \{2, \ldots, n\}$. For each $k \in F$, a clustering algorithm divides the data set in k clusters, and the average silhouette $\text{avg}(s(i))_k$ is computed for all k clusters. To conclude, the final average $\sum_{k=2}^{n} \text{media}(s(i))_k$ is computed and selected as the best number of clusters for the data set.

2.1.2. Partitioning around medoids (PAM)

The clustering algorithm selected was the PAM [24], an extension of the K-means algorithm. K-means is an antique clustering algorithm which is still very popular due to its speed, efficiency and simplicity [22]. This method searches for k centers within the data set which minimizes the total sum of the squared distances between each sample and its nearest center. The K-means is executed in the following steps:

1. Randomly choose k samples from the data set as the initial centers $C = \{c_1, c_2, \ldots c_k\}$.
2. For each $i \in \{1,\ldots,k\}$, set the cluster C_i as the subset of samples which are nearest to c_i than to any other center c_j for all $j \neq i$.
3. For each $i \in \{1,\ldots,k\}$, update c_i to be the center of mass of all samples in C_i.
4. Repeat Steps 2 and 3 until no changes are observed in C or a maximum number of iterations is reached.

Although K-means is currently the best-known and most used clustering algorithm, it has the key limitation of working with numerical values only, which makes its application on our mixed data set unfeasible. We opted for the use of the PAM algorithm, which implements the K-medoids [23]clustering algorithm.

K-medoids is similar to K-means, but more robust and less sensitive to outliers. The k-medoids algorithm is based on finding k representative samples within the data set, called medoids, and the k clusters will be constructed through the association of each sample to its nearest representative object. K-medoids can also receive a dissimilarity matrix instead of the data, a matrix which

Table 1
Type and description of the descriptive variables, which correspond to information obtained on 476 family farmers from six country rural towns in Ceará state.

Identificação	Tipo	Descrição
Municipality	Categoric	City where the family resides in
F_1	Binary	Household has saving account
F_2	Binary	Household received credit in the past
F_4	Binary	Household received remittances
F_6	Binary	Household never received Bolsa Família
H_1	Binary	Household has access to health services
H_5	Binary	School enrollment
H_5a	Binary	Household has kids in school age
N_1	Binary	Household own their own land
N_3	Binary	Household own more land than they cultivate
N_4	Nominal	Dependability of irrigation source
N_5	Nominal	Household have made land improvements
N_7	Binary	Household has running water
M_3	Binary	Household owns a car or motorcycle
M_4	Binary	Household uses a tractor
S_2c	Binay	Household provided support to other families
S_2d	Binary	Household received support from other families
AS_1	Ratio	Number of plots of agricultural land
AS_5	Binary	Household has total subsistence crop loss
P_1	Binary	Household participated in “Hora de Plantar” government seed distribution program
P_2	Binary	Household has knowledge of PAA (government food purchasing program)
P_4	Binary	Household has knowledge of crop insurance program
P_5	Binary	Household participated in crop insurance
P_6	Binary	Household used loaned farm equipment
P_7	Binary	Household participated in PAA
FS_3	Binary	Household has insufficient quantity of food
FS_4	Binary	Household has food of insufficient quality
MG_1	Binary	Household has an individual that migrated in the past and has returned
MG_4	Binary	Household has an individual that is a current migrant
MG_6	Nominal	Types of migrants in the household
AG_1	Binary	Household planted corn or beans for subsistence
AG_2	Binary	Household planted other subsistence crops
AG_3	Binary	Household planted horticulture
AG_4	Binary	Household planted cash crops
AG_5	Binary	Household planted fruit crops
F_3	Numeric	Per capita annual income
F_3a	Numeric	Per capita annual income, inflation adjusted
F_5	Numeric	Per capita livestock assets
H_2	Ratio	Dependency ratio
H_3	Ratio	Percentage of adults older than 17 that have finished high school
N_2	Numeric	Quantity of land cultivated in hectares
M_1	Ordinal	Number of pieces of farm equipment owned
M_2	Ratio	Asset index
M_2a	Ratio	Household asset index
M_2b	Ratio	Age productivity asset index
M_5	Numeric	Number of agricultural technologies used
S_1	Numeric	Number of social groups that the household belongs to
S_2	Ratio	Frequency of given and received support
S_2a	Ratio	Frequency of support given
S_2b	Ratio	Frequency of support received
LS_1	Numeric	Number of sources of income per capita
LS_2a	Ratio	Percentage of total income from climate sensitive sources
LS_2b	Ratio	Percentage of total income from climate neutral sources
LS_2c	Ratio	Percentage of total income from Bolsa Família cash transfer program
LS_2d	Ratio	Percentage of total income from social security
LS_3	Ordinal	Drought impacts on household relative to other households in the area
AS_2	Ordinal	Soil quality index
AS_3	Ratio	Number of crop types
FS_1	Ordinal	Food security index
FS_2	Ratio	Acess index
MG_2	Numeric	Number of migration episodes in the past per household
MG_3	Numeric	Maximum number of individual migration episodes
MG_5	Numeric	Number of current migrants
MG_7a	Numeric	Sum of individuals in household that migrated to city seat
MG_7b	Numeric	Sum of individuals in household that migrated to another city
MG_7c	Numeric	Sum of individuals in household that migrated to Fortaleza (capital of Ceará)
MG_7d	Numeric	Sum of individuals in household that migrated to another state
MG_7e	Numeric	Sum of individuals in household that migrated to another country
R_1	Numeric	Annual household income from climate sensitive sources
R_2	Numeric	Annual household income from social security
R_3	Numeric	Annual household income from Bolsa Família
R_4	Numeric	Annual household income from climate neutral sources
R_1a	Numeric	Annual household income from climate sensitive sources, inflation adjusted
R_2a	Numeric	Annual household income from social security, inflation adjusted
R_3a	Numeric	Annual household income from Bolsa Família, inflation adjusted
R_4a	Numeric	Annual household income from climate neutral sources, inflation adjusted

Table 2
Best features determined by the CFS method.

Feature	Type	Description
F_2	Categorical	Household received credit in the past
F_3	Numeric	Per capita annual income
F_6	Categorical	Household never received Bolsa Família
H_1	Categorical	Household has health access
S_2a	Numeric	Frequency of support given
S_2c	Categorical	Household gave support to other families
LS_2C	Numeric	Percentage of total income from Bolsa Família cash transfer program
LS_2D	Numeric	Percentage of total income from social security
P_5	Categorical	Household participated in crop insurance
FS_1	Numeric	Food security index
FS_2	Numeric	Access index
FS_4	Categorical	Household has food of insufficient quality
R_2	Numeric	Annual household income from social security
R_2A	Numeric	Annual household income from social security, 2012 inflation adjusted
R_3A	Numeric	Annual household income from Bolsa Família, 2012 inflation adjusted
AG_1	Categorical	Household planted corn or beans for subsistence

stores pairwise distances between the samples. The medoid of a cluster will be the sample from this cluster such that the average dissimilarity between this sample and the other samples in this same cluster is as small as possible, instead of the squared sum of Euclidian distances computed by the K-means. The algorithm is executed in two steps, BUILD and SWAP. In BUILD phase, an initial clustering is performed by the successive selection of k representative samples. The first sample selected must present the least average dissimilarity possible from the other samples and, therefore, should be located at the center of mass of the data set. The next $k-1$ samples are selected in order to decrease the objective function value, following the steps:

1. Consider i as an unselected sample.
2. Consider j an unselected sample. Compute the difference between D_j (dissimilarity between j and the nearest selected sample) and $d(j,i)$ (dissimilarity between j and i).
3. If the computed difference is positive, compute the contribution of j to the decision if i should be selected, given by $C_{ji} = \max(D_j - d(j,i), 0)$.
4. Compute the total gain obtained if i is selected given by $\sum_j C_{ji}$.
5. Select the unselected sample i which satisfies $\max_i \sum_j C_{ji}$.

The SWAP phase tries to optimize the set of selected representative samples. Consider all pairs (i,h) of selected sample i and unselected sample h. In order to decide if i should be swaped with h, the following equations and steps are performed:

1. Consider an unselected sample j and compute its contribution C_{jih} to the swap:
   a. If j is more distant from i and h than from the other representative samples, C_{jih} is zero.
   b. If j is not more distant from i than from the other representative samples,$d(j,i) = D_j$;
   i. If j is nearest to h than to the second nearest representative sample, $(d(j,i) < E_j$, where E_j is the dissimilarity between j and the second nearest representative sample), the contribution of j to the swap between i and h is given by $C_{jih} = d(j,h) - d(j,i)$.
   ii. If j is equally distant from h than from the second nearest representative sample, $(d(j,i) \geqslant E_j)$, then the contribution of i to the swap is $C_{jih} = E_j - D_j$.
   iii. If j is more distant from i than from at least one of the other representative samples, but nearest to h than from any other representative sample, the contribution of j to the swap is $C_{jih} = d(j,h) - D_j$.
2. Compute the final result of the swap by adding the contributions $T_{ih} = \sum_j C_{jih}$.
3. Select the pair (i,h) which satisfies the function $\min_{i,h} T_{ih}$.
4. If the minimum value of T_{ih} is negative, swap and return to step 1. Otherwise, the swap is considered undesirable and the algorithm ends.

2.1.3. Dissimilarity matrix calculation

The dissimilarity matrix which PAM algorithm receives as input parameter can be computed using the Daisy function detailed by [24]. This function accepts mixed data types, including numeric, categoric, ordinal, symmetric and asymmetric binary values, which is useful for the data set we analyze in this study. In order to handle mixed variables, Daisy uses the Gower dissimilarity coefficient [14]. Each variable is normalized by dividing each value by the range of values of the corresponding variable, after the subtraction of the minimum value. Thereafter, the variable will be scaled to [0,1]. Null values are discarded from the calculation.

The Gower coefficient computes the distance between two samples i and j as the weighted average of the contributions of each variable, given by the equation:

$$d_{ij} = d(i,j) = \frac{\sum_{k=1}^{p} w_k \delta_{ij}^k d_{ij}^k}{\sum_{k=1}^{p} w_k \delta_{ij}^k}$$

where d_{ij} is the weighted average of d_{ij}^k, $w_k \delta_{ij}^k$ are the corresponding weights, δ_{ij}^k is 0 or 1, and d_{ij}^k is the kth contribution of the variable for the total distance. The weight δ_{ij}^k is 0 when variable k presents a null value for i or j, or 1 otherwise. The contribution d_{ij}^k of a categoric variable for the total distance is 0 if both values are equal 1, or 0 otherwise. For the other types of variable, the contribution is the difference between the two values divided by the range of values of the variable.

2.2. Data classification

In addition to cluster analysis, data mining also offers techniques to perform predictive analyses called classification. Classification models implement supervised learning, in which a certain information represented by a class label of an unknown sample can be predicted based on previous observation of labeled samples. These labeled samples are also called training set. The final product of supervised learning is a classifier that is mathematically described by a function $\hat{f} : X \to \{true, false\}$ which uses a subset of D and $\hat{f}(x) \cong f(x)$, being D a set of labeled samples and $D = \{(x,y) | x \in Sey = f(x)\}$ for a labeling function $f : X \to \{true, false\}$ [20].

After clustering the data, we developed classification models capable of predicting in which cluster an arbitrary family would be associated to. We selected three of the most popular classification models in the data classification literature: artificial neural networks, support vector machines and decision trees.

2.2.1. Decision trees

Decision trees [31] refers to one of the oldest classification models. Each variable of the data set is individually questioned, and all the questions and answers can be arranged in a hierarchical structure called a tree. Each node of the tree refers to a variable, and each edge originated in a node represents a value, or a range of

values, for the variable represented by the node. Leaf nodes store the class labels, the final point of the classification. Whenever a new test sample is analyzed, each one of the variables are questioned, following a preexisting path along the tree until a leaf node is reached and its corresponding class label is set as the class label predicted for the analyzed sample. The recursive Hunt algorithm is one of the most commonly used algorithms for building decision trees. The idea behing it is that the nodes of the tree should be arranged in order to maximize the information gain, i.e., the questioned variables should be capable of dividing the data set in pure partitions, with larger frequencies of a determined class.

Another advantage of decision trees is the ease of visualization and interpretation. The decision rules exposed on the paths allow us to generate hypotheses regarding the individual influence of each variable in the classification process. Therefore, although they are used primarily as classification models, in this study we employ decision trees in order to interpret the clustering results and visualize which variables are the most relevant to the separation of the clusters [25,35].

2.2.2. Support vector machines

Support vector machine (SVM) is a popular classification model in the recent data mining literature due to its efficiency and empirical success. Data sets usually have more than one decision boundary capable of dividing the data in the n-dimensional space, and the goal of SVM is to compute the hyperplane with the largest margin possible to act as a decision boundary [35]. When applied to data that is linearly separable, the decision boundary has a linear representation given by the equation:

$$w \cdot x + b = 0$$

where x refers to the variable set of an arbitrary sample e, and w is a set of weights whose linear combination gives the class label y for e. The width of the decision boundary margin is given by the distance between the two parallel hyperplanes (support vectors) which touch the samples from each class that is nearest to the decision boundary, and is defined as [7,10]:

$$d = \frac{2}{||w||}$$

Finally, the goal of SVM is to find the decision boundary with the largest margin possible, which means minimizing the following objective function:

$$\min_w \frac{||w||^2}{2}$$

When handling data that is not linearly separable, SVM projects the data from its original n-dimensional space into a new space where the samples can be separated by a linear decision boundary. This process is achieved with the aid of kernel functions $K(x,y)$. Kernel functions express the similarity between two samples in the new transformed dimensional space according to their dot product. In this study, we use the radial basis kernel function (RBF) to transform the data, given by the equation:

$$K(x,y) = \exp(-\gamma||x-y||^2)$$

2.2.3. Artificial neural networks

Artificial neural networks are inspired in the cognitive system and neurological functions of the human brain, simulating its neuron and links. A neuron has a filament called axon which connects to another neuron through dendrites, and the connection point is called synapse. Similarly, a classification model based on artificial neural networks is composed of interconnected nodes [35]. In this study, we use a classification model of this family called Multilayer Perceptron (MLP).

A perceptron is a structure composed of n input nodes that receive the values from the feature variables which describe the samples. Each node from this initial layer is linked to the output node through an edge that is associated to a weight value w. The output value is determined by the equation:

$$f = \left(\sum_{k=1}^{n} w_k x_k\right) - t$$

The class label is 1 if $f > 0$ and 0 otherwise. The model's performance can be improved by systematic adjustments in the weigth values $w_1, w_2, \ldots, w_n$. The MLP has hidden layers between the initial layer and output layer which works with backpropagation: the prediction errors obtained during the training phase are backpropagated to the previous layers, and this error value is used to adjust the weights in each edge. The training phase with the implementation of the backpropagation is summarized in the following steps [13]:

1. Initialize the neural network with the weight values in each edge
2. Read the first input sample
3. Propagate the feature values of the given samples through the neural network until an output value is obtained
4. Compute the error value by comparing the output value obtained with the expected output value
5. Propagate the error back through the network
6. Adjust the weigth values in order to minimize the overall classification error
7. Repeat steps 2–7 for new training samples, until the overall error is minimized

2.2.4. Evaluating the performance of classification models

The evaluation of classification models is performed by using a set of training samples to train the model and a set of test samples formed by unknown samples to test it. In the k-fold cross validation method, the data set D is randomly divided into k mutually exclusive subsets $D_1, D_2, \ldots, D_k$ (folds) of equal or similar size. The classification model is trained and tested k times, and in each iteration $t \in \{1, 2, \ldots, k\}$, the folds $D - D_t$ are used to train the model and the remaining fold D_t is used to test it [12]. Given the classification results obtained during each test phase, the accuracy of the model is computed as:

$$Accuracy(\%) = \frac{c}{n} \times 100\%$$

where c is the number of test samples which were correctly classified, and n is the total number of test samples. The final accuracy of the model estimated by the k-fold cross validation is the average of all accuracies obtained in each iteration.

2.3. Feature selection

Feature selection is a data mining process which aims for the removal of feature variables which are considered unimportant of the classification. In general, we say that a feature is considered important for the analysis if it is relevant and irredundant [9]. A variable is considered redundant when it presents high dependence to the other variables and the information contained in it can be expressed by a fewer number of these variables. A variable is considered irrelevant when the information it contains does not contribute to generate hypotheses regarding the samples with relation to their class labels. The removal of such variables can

improve the accuracy of classification models, reduce outliers and simplify the analysis [11,18].

2.3.1. Correlation Based Feature Selection

CFS (Correlation Based Feature Selection) is a feature selection method proposed by Hall [19] which employs a heuristic based on correlations to estimate the importance of a subset of features. This method generates subsets with different combinations of features, and each subset is evaluated by the following Pearson correlation equation:

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

being k the total number of features, M_s is the merit of a feature subset S, $\overline{r_{cf}}$ is the average correlation between the features and the class label, and $\overline{r_{ff}}$ is the average intercorrelation between the features. Irrelevant features present low value for $\overline{r_{cf}}$ and redundant features present high value for $\overline{r_{ff}}$, and both cases lead to a minimization of the M_s value. Finally, the feature subset which achieved the highest merit is selected as the best feature subset.

3. Results

3.1. Groups identified

The first step of the analysis was estimating the empirically best number of clusters within the data set. We determined the average width of the silhouette for 2–50 clusters, and we found that the analyzed families could be divided into a best number of 2 clusters. The PAM clustering algorithm was applied, and the data was divided into 2 clusters of 224 and 252 samples, respectively. A t-SNE (t-Distributed Stochastic Neighbor Embedding) graph for the partition is presented in Fig. 2, allowing the visualization of the formed clusters in a bidimensional plot. The frequency of families from each city in the clusters is shown in Figs. 3 and 4.

In Fig. 3, we can see that families from Barbalha and Parambu cities are more frequent in Cluster 1, corresponding to 21% and 20% respectively of the total families associated to this cluster. This percentage varies between 10% and 17% for the other cities. Families from Guaraciaba do Norte are less frequent in this cluster, accounting for 10% of the families. This scenario is reversed for Cluster 2, presented in Fig. 4. This cluster has more families from Guaraciaba do Norte (23%) and fewer families from Barbalha and Parambu (13% both). Figs. 5–10 bring another visualization of this result, presenting the frequency of the two clusters in each city. As a matter of fact, we can see that families from Boa Viagem (Fig. 6)

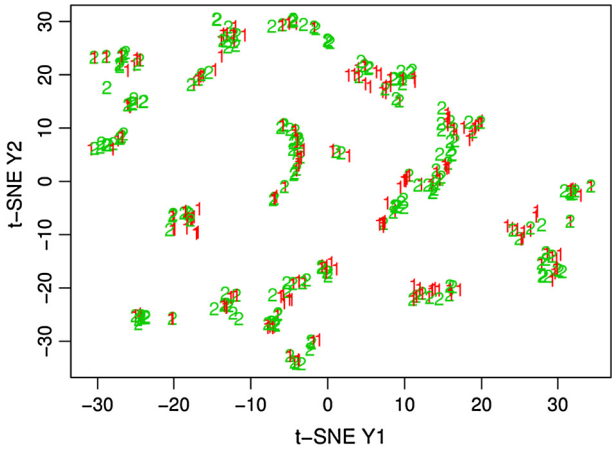

Fig. 2. The t-SNE bidimensional graph for the found clusters.

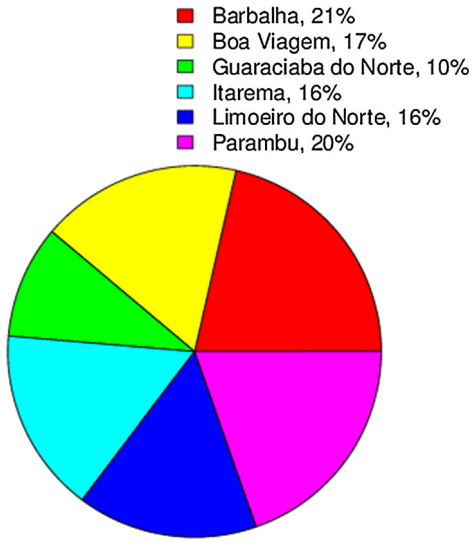

Fig. 3. Frequency of cities in Cluster 1. 21% Barbalha, 17% Boa Viagem, 10% Guaraciaba do Norte, 16% Itarema, 16% Limoeiro do Norte, 20% Parambu.

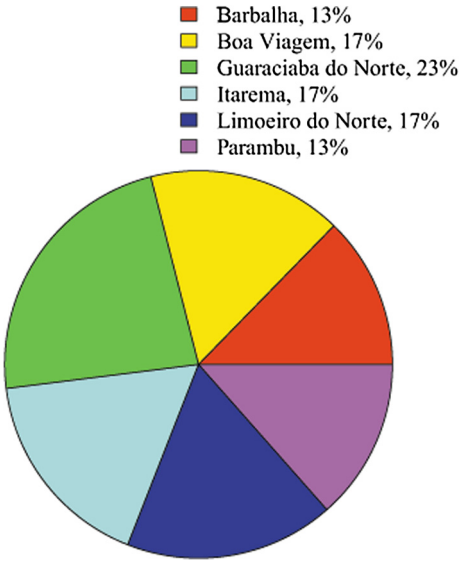

Fig. 4. Frequency of cities in Cluster 2. 13% Barbalha, 16% Boa Viagem, 23% Guaraciaba do Norte, 17% Itarema, 17% Limoeiro do Norte, 13% Parambu.

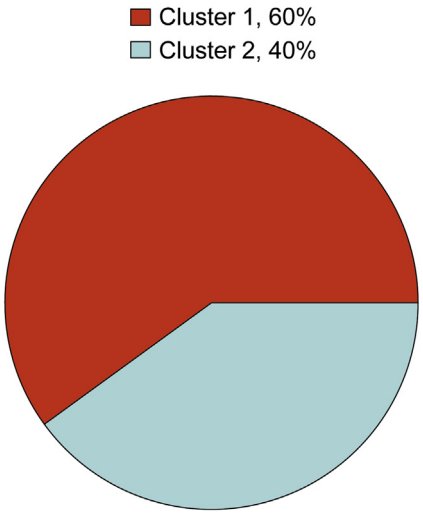

Fig. 5. Barbalha. 60% Cluster 1, 40% Cluster 2.

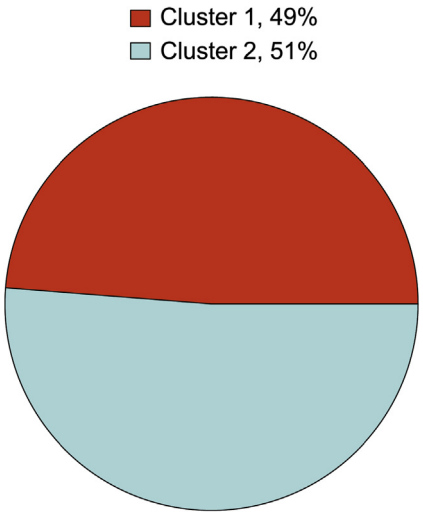

Fig. 6. Boa Viagem. 49% Cluster 1, 51% Cluster 2.

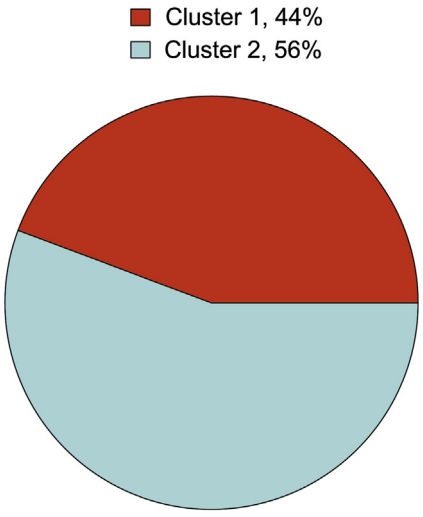

Fig. 9. Limoeiro do Norte. 44% Cluster 1, 56% Cluster 2.

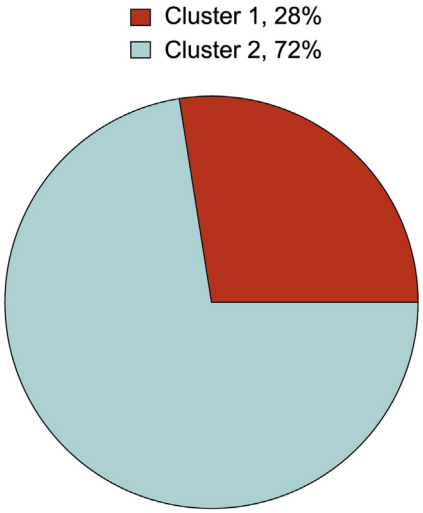

Fig. 7. Guaraciaba do Norte. 28% Cluster 1, 72% Cluster 2.

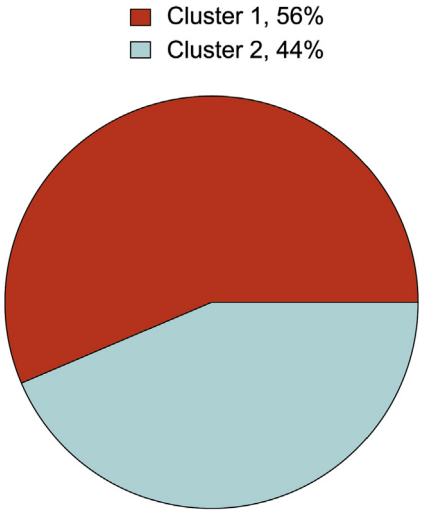

Fig. 10. Parambu. 56% Cluster 1, 44% Cluster 2.

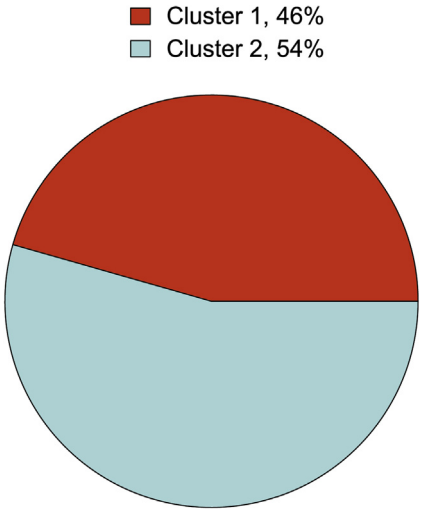

Fig. 8. Itarema. 46% Cluster 1, 54% Cluster 2.

and Itarema (Fig. 7) are equally distributed in both clusters. Families from Guaraciaba do Norte were mostly associated to Cluster 2 (72%), and families from Barbalha were mostly associated to Cluster 1 (60%).

3.2. Feature selection and classification

After determination of the clusters, we considered the cluster value of each sample as class label and proceeded to develop classification models capable of predicting in which cluster a new, unknown family, would be associated to. We also applied the CFS algorithm to determine the most relevant features which differentiate the families of the two clusters. The best feature subset estimated by the CFS is shown in Table 1.

The prediction results obtained by the classification model before and after feature selection are presented in Table 3. It is visible that both SVM and MLP models achieved best accuracy when all feature variables are used for training. The classification model which achieved the best performance was the MLP with 93.48% prediction accuracy when all variables are used.

Table 3
Accuracy values achieved by the classification models before and after feature selection is performed.

Model	Before CFS	After CFS
SVM	92.22%	86.76%
MLP	93.48%	85.50%
Decision trees	81.51%	85.50%

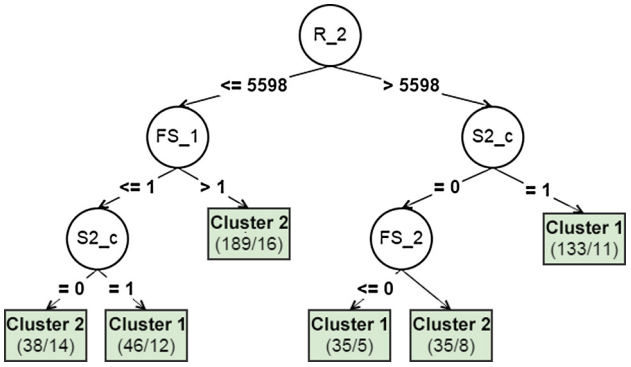

Fig. 11. Decision tree built for the analyzed data set, considering only the most important features selected by the CFS method and the cluster value as class label.

3.3. Interpretation and discussion

We used Weka software (version 3.6) to build a decision tree considering the cluster value of the samples as class label and only the most important features determined by the CFS as input variables. The algorithm employed was the C4.5, which yielded 85.5% prediction accuracy. The structure of the decision tree achieved is shown in Fig. 11.

The first feature selected by the C4.5 to divide the samples and which present the highest information gain is R_2 (annual household income from social security), which means that we can organize the results in two cases: when the families have annual household income from social security relatively higher than the others (> R$5598), and otherwise.

Considering the families which have annual income from social security higher than R$5598, we observe the following patterns:

- Cluster 1: in this cluster, the families with higher income from social security give support to others more frequently, otherwise they are characterized by a minimum access index.
- Cluster 2: in this cluster, the families with higher income do not give support to others frequently, and have higher access index than the families with higher income associated to Cluster 1.

Considering the families which have annual income from social security less than R$5598, we observe the following patterns:

- Cluster 1: in this cluster, families with lower income from social security have the minimum value of food security index (1), but they give support to other families more frequently.
- Cluster 2: in this cluster, families with lower income from social security have a higher food security index, and families with low food security index do not give support to other families frequently.

An individual analysis of the variables and decision rules presented in the tree's structure lead us to the following conclusions:

- Annual household income from social security (R_2): Considering the prediction error rate, the tree shows that a total of 160 families (71%) associated to Cluster 1 have income from social security higher than R$5598, while only 43 families (17%) of the families in Cluster 2 have income from social security also higher than this value. Therefore, the majority of families associated to Cluster 1 have income from social security higher than families associated to Cluster 2.
- Food security index (FS_1): This variable considers whenever the family has income from social security lower than R$5598. Given the 64 families from Cluster 1 which have this income, 48 families (21% of the total of families of Cluster 1) are characterized by minimum food security index. On the other hand, in Cluster 2, 173 families (69% of the total of families associated to Cluster 2) have income from social security lower than R$5598 and food security index higher than the minimum. Therefore, in the few cases where a family from Cluster 1 has relatively low income from social security, it is likely that its food security index is also minimum. Moreover, the majority of families from Cluster 2 have income from social security lower than the families from Cluster 1, however, they present higher food security index overall.
- Household gave support to other families (S2_c): This variable is always considered to determine if a family belongs to Cluster 1. Around 70% of the families associated to Cluster 1 do give support to other families. Considering the remaining 30% families that do not give support, 13% have relatively high income from social security, but are characterized by the minimum access index (FS_2), and 11% have relatively low income from social security and minimum food security index. For Cluster 2, we can see that, considering the 43 families who have income from social security higher than R$5598, 32 of them did not give support the other families. Therefore, we can conclude that families associated to Cluster 1 usually gave support to other families.

Overall, the majority of families from Cluster 1 receive income from social security higher than the families in Cluster 2, and gave support to other families even in the rare cases where they show lower income from social security and minimum food security index. When families of this cluster did not gave support, they present a minimum access index. As discussed previously, the majority of families from Barbalha (60% of the analyzed families) and Parambu (56%) belong to this cluster.

On the other hand, families from Cluster 2 received lower income from social security, but they had a higher access index when they received lower income, and present higher food security index when the income from social security was low. As discussed previously, most families from Boa Viagem, Guaraciaba do Norte, Itarema and Limoeiro do Norte are associated to this cluster – approximately 51%, 72%, 54% and 56%, respectively.

4. Conclusion

We presented a data mining study and cluster analysis of social data obtained from small producers and family farmers from six municipalities in Ceará state, Northeastern Brazil. The social data were obtained through research conducted directly with the families, which involved personal questions regarding demography, economics, agricultural production, and food security. The answers obtained were used as feature variables for our analysis. In order to work with this challenging data, we employed a methodology capable of handling variables of mixed types and null values. We estimated the best number of clusters within the data with the Silhouette technique, and the PAM clustering algorithm was employed in order to partition the data in the clusters. With the aid of the C4.5 decision tree, we observed that, overall, families from Cluster 1 received higher income from social security than the families associated to Cluster 2, and gave more support to other

families, even in the rare cases where the families presented lower income from social security and a in minimum value of food security. When families of the cluster did not give support to others, they present a minimum access index. Most of the analyzed families from Barbalha and Parambu belong to this cluster. On the other hand, families from Cluster 2 receive less income from social security, but they have a higher access index when the income from social security is higher. These families also have a higher food security index, even when the income from social security is low. Most of the analyzed families from Boa Viagem, Guaraciaba do Norte, Itarema and Limoeiro do Norte are associated with this cluster.

The proposed methodology presented promising results, however, there are a few issues that could be addressed in future works.

First, we worked with a relatively small data set, and the analysis of a larger number of families could help us uncover more trends regarding their profile. Moreover, we investigated families from Ceará state only, and it would be interesting to extend our analysis to families from other states as well. Piauí, Rio Grande do Norte, Paraíba and Pernambuco are states bordering Ceará and located in the Northeast region as well, sharing very similar geological, climatic and demographic characteristics, and families from these states could be inserted into the analysis in order for us to investigate demographic, economic and social profiles of small producers and family farmers of the Northeast region as a whole instead of one state only.

Also, working with categorical values is one of the major challenges for data mining researches. We expect, for future works, to improve the processability of our data set, either by standardizing our variables to numeric values only, changing the data collection method (i.e., replace the common surveys) or by employing forthcoming clustering and classification techniques in the data mining literature that can handle categorical variables.

Declarations of interest

None.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] A. Ahmad, L. Dey, A k-mean clustering algorithm for mixed numeric and categorical data, Data Knowl. Eng. 63 (2007) 503–527.
[2] G. Barbier, H. Liu, Data Mining in Social Media, in: Social Network Data Analytics, Springer, US, Boston, MA, 2011, pp. 327–352.
[3] R.M. Barbosa, B.L. Batista, C.V. Barião, R.M. Varrique, V.A. Coelho, A.D. Campiglia, F. Barbosa, A simple and practical control of the authenticity of organic sugarcane samples based on the use of machine-learning algorithms and trace elements determination by inductively coupled plasma mass spectrometry, Food Chem. 184 (2015) 154–159.
[4] R.M. Barbosa, B.L. Batista, R.M. Varrique, V.A. Coelho, A.D. Campiglia, F. Barbosa, The use of advanced chemometric techniques and trace element levels for controlling the authenticity of organic coffee, Food Res. Int. 61 (2014) 246–251.
[5] C.M. Bishop, Pattern Recognition and Machine Learning, Springer – Verlag New York, 2006.
[6] R.P. Bunker, F. Thabtah, A Machine Learning Framework for Sport Result Prediction, Appl. Comput, Informatics. 15 (2019) 27–33.
[7] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (1998) 121–167.
[8] J.N.B. Campos, Paradigms and public policies on drought in Northeast Brazil: a historical perspective, Environ. Manage. 55 (2015) 1052–1063.
[9] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (2014) 16–28.
[10] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[11] M. Dash, H. Liu, Feature selection for classification, Intell. Data Anal. 1 (1997) 131–156.
[12] R.O. Duda, P.E. Hart (E. Peter, D.G. Stork). Pattern classification. Wiley, 2001.
[13] M. Gardner, S. Dorling, Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, Atmos. Environ. 32 (1998) 2627–2636.
[14] J.C. Gower, A general coefficient of similarity and some of its properties, Biometrics 27 (1971) 857.
[15] Guanziroli, C.E., 2001. Agricultura familiar e reforma agrária no século XXI. Garamond.
[16] C.E. Guanziroli, A.M. Buainain, A. Di Sabbato, Dez anos de evolução da agricultura familiar no Brasil: (1996 e 2006), Rev. Econ. e Sociol. Rural 50 (2012) 351–370.
[17] P. Gundecha, H. Liu, Mining social media: a brief introduction, in: Tutorials in Operations Research. INFORMS, 2012, pp. 1–17.
[18] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[19] M.A. Hall, Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning, in: Pat Langley (Ed.), Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000). Morgan Kaufmann Publishers Inc., Stanford, CA, USA, 2000, pp. 359–366.
[20] L. Hamel, Knowledge Discovery with Support Vector Machines, Wiley-Interscience, 2009.
[21] Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Min. Knowl. Discov. 2 (283–304) (1998) 1.
[22] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recogn. Lett. (2010).
[23] L. Kaufman, P. Rousseeuw, Clustering by means of medoids, in: Y. Dodge (Ed.), Statistical Data Analysis Based on the L1-Norm and Related Methods, 1987, pp. 405–416.
[24] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, 2005.
[25] D.M. Khan, N. Mohamudally, An integration of K-means and decision tree (ID3) towards a more efficient data mining algorithm, J. Comput. 3 (2011) 76–82.
[26] C. Maione, E.S. De Paula, M. Gallimberti, B.L. Batista, A.D. Campiglia, F. Barbosa, R.M. Barbosa, Comparative study of data mining techniques for the authentication of organic grape juice based on ICP-MS analysis, Expert Syst. Appl. 49 (2016) 60–73.
[27] C. Maione, B. Lemos, A. Dobal, F. Barbosa, R. Melgaço, Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry, Comput. Electron. Agric. 121 (2016) 101–107.
[28] A. Mucherino, P. Papajorgji, P.M. Pardalos, A survey of data mining techniques applied to agriculture, Oper. Res. 9 (2009) 121–140.
[29] M.R. Naroui Rad, S. Koohkan, H.R. Fanaei, M.R. Pahlavan Rad, Application of Artificial Neural Networks to predict the final fruit weight and random forest to select important variables in native population of melon (*Cucumis melo* L.), Sci. Hortic. (Amsterdam) 181 (2015) 108–112.
[30] E.W.T. Ngai, L. Xiu, D.C.K. Chau, Application of data mining techniques in customer relationship management: a literature review and classification, Expert Syst. Appl. 36 (2009) 2592–2602.
[31] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.
[32] M.V.J. Reddy, B. Kavitha, Clustering the mixed numerical and categorical dataset using similarity weight and filter method, Int. J. Database Theory Appl. (2012) 5.
[33] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[34] I. Scoones, Livelihoods perspectives and rural development, J. Peasant Stud. 36 (2009) 171–196.
[35] P.-N. Tan, M. Steinbach, V. Kumar, Introduction to data mining. Pearson Addison Wesley, 2005.