61

Presentation date: September, 2021
Date of acceptance: November, 2021
Publication date: December, 2021

K-MEANS

BASED METHOD FOR HANDLING UNLABELED DATA

MÉTODO BASADO EN K-MEANS PARA EL TRATAMIENTO DE DATOS NO ETIQUETADOS

Sharon Diznarda Álvarez Gómez[1]
E-mail: dirfinanciera@uniandes.edu.ec
ORCID: https://orcid.org/0000-0003-3213-9034
Silvio Amable Machuca Vivar[2]
E-mail: c.investigacionstd@uniandes.edu.ec
ORCID: https://orcid.org/0000-0002-4681-3045
Paulina Elizabeth Salas Medina[3]
E-mail: ua.paulinasalas@uniandes.edu.ec
ORCID: https://orcid.org/0000-0001-6573-533X
[1] Universidad Regional Autónoma de Los Andes. Ecuador.

Suggested citation (APA, 7th edition)

Álvarez Gómez, S. D., Machuca Vivar, S. A., & Salas Medina, P. E. (2021). K-means based method for handling unlabeled data. *Revista Universidad y Sociedad*, 13(S3), 452-458.

ABSTRACT

From the development achieved by the current information society, incalculable volumes of data are generated. The exponential growth of information significantly supports people's decision making in their daily activities. In Ecuador there are many institutions that store the data of their processes, the tourism sector representing an example of this. However, the data generated exceeds the power of analysis and processing of human beings, sometimes relevant information is presented that is not visible to people. The present investigation proposes a solution to the described problem starting from the development of a method for the treatment of unlabeled data.The proposed method is based on the unsupervised k-means algorithm. The proposal has been implemented from the stored data set of the tourism sector in the City of Riobamba.

Keywords: Machine learning, data mining, roughsets, entropy, information gain.

RESUMEN

A partir del desarrollo alcanzado por la actual sociedad de la información, se generan volúmenes incalculables de datos. El crecimiento exponencial de la información apoya significativamente la toma de decisiones de las personas en sus actividades cotidianas. En el Ecuador existen muchas instituciones que almacenan los datos de sus procesos, el sector turístico representa un ejemplo de ello. Sin embargo, los datos generados superan el poder de análisis y procesamiento del ser humano, en ocasiones se presenta información relevante que no es visible para las personas. La presente investigación propone una solución al problema descrito a partir del desarrollo de un método para el tratamiento de datos no etiquetados, basado en el algoritmo no supervisado de k-means. La propuesta ha sido implementada a partir del conjunto de datos almacenados del sector turístico de la ciudad de Riobamba.

Palabras clave: Aprendizaje automático, minería de datos, roughsets, entropía, ganancia de información.

INTRODUCTION

Tourism represents an important source of income in Ecuador's internal economy.Each region of the country has attractions that make it unique as a tourist destination. The city of Riobamba in Ecuador is characterized by representing a very attractive tourist area, it is a city with great cultural heritage that attracts exquisite vacationers (Brachtl et al. 2009)

Tourism management itself generates high demands for products and services that include a wide range of different activities such as: transportation to destinations, accommodation, supply, shopping, travel agency services, inbound and outbound tourism operators, among others (Sierra, 2016). Without a doubt, tourism represents a fundamental source of income that generates a large amount of data.

From the different operations that are carried out in the City of Riobamba, there are stored historical data of the different operations that are carried out in tourism management (García et al. 2017). However, the existing data is not properly labeled, which makes it impossible to obtain objective information that contributes to decision-making for the tourism sector. (Ricardo et al. 2019).

Problems of this nature have been addressed in the scientific literature from data mining techniques for the cleaning, transformation and treatment of unlabeled data (Baalaji & Khanaa, 2020). The present investigation defines as objective to develop a method based on k-means for the treatment of unlabeled data.

Preliminaries

This section introduces an approximation of the main theoretical references that support the research proposal.It begins with a characterization of machine learning. The fundamental elements on the rough sets are presented. Some criteria for comparing k-means algorithms are presented. The section continues with the significant elements associated with entropy and information gain. Finally, the used k-means algorithm is described.

Machine learning

Machine learning introduces a new paradigm that refers to the study of computational algorithms that nurture its operation to automatically incorporate experiences to improve its operation. Machine learning systems simulate the processes humans perform when performing a task.

A machine learning process needs to train a model by applying learning techniques. For the training process, data is provided that the machine will use to learn this procedure (Arnaiz et al. 2020). This type of learning has been used in data mining applications with the aim of discovering rules and patterns in large data sets and filtering information (Shokri et al. 2017).

As for the classification of machine learning techniques can be divided:

- Supervised or predictive learning: where the objective is to learn to map from X inputs to Y outputs, given a labeled set of N input-output pairs; this set is called Training set.
- Unsupervised or descriptive learning: aims to find interesting patterns in the N entries.
- Reinforcement of learning: it is used to know how it acts or behaves when certain occasional signs of reward or punishment are given.

Rough sets

The Roughsets (RS) are based on the assumption that each object x in the universe of discourse U has associated certain information that represents data and knowledge. It is expressed through attributes that describe the object. Among the advantages of RS for data analysis are (Fraser & Yu, 2021):

- It is based on the original data and does not require external information, so there is no need to make any assumptions about the data.
- It allows the analysis of qualitative and quantitative traits.

Then a rough set is formalized.

Let U be a finite universe. Let R be an equivalence relation defined in U, which partitions U. (***U, R***) is a collection of all equivalence classes, called the approximation space. Let w_1, w_2, w_3, w_n elements of the approximation space (***U, R***). This connection is known as the knowledge base. Then, for any subset B of U, the top approximation $\overline{B}$ and the bottom approximation $\underline{B}$ are defined as (Machado, 2018).

The ordered pair $(\overline{B}, \underline{B})$ is called an approximate set and it also has:

$POS(B) = \underline{B} \rightarrow$ he is certainly a member of X.

$NEG(B) = U - \overline{B} \rightarrow$ he's certainly not a member of X.

$BR(B) = \overline{B} - \underline{B} \rightarrow$ possibly a member of X.

Where:

POS(B): refers to the positive region of B,

NEG(B): refers to the negative region of B,

BR(B): refers to the border region of B.

Entropy and information gain

The entropy in a data source represents the magnitude that measures the information provided about the data source. Entropy provides information about a specific data source or fact (Crevecoeur, 2019):

Definition

Given two classes P and N in a sample space S, where:

$$S=P \cup N \quad (1)$$

Where the cardinality is given by:

$$|P|=p \text{ and } |N|=n \quad (2)$$

Entropy refers to the amount of information necessary to decide whether a sample of S belongs to P or to N and is defined as (Isler et al. 2016; Carrillo et al. 2018).

$$E(S) = \frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \quad (3)$$

When selecting an attribute b the sample space is divided into child subsets of b, the way to determine how much information an attribute b contributes in a total set of attributes A, is given by (Sadri et al. 2017).

$$Input(b) = E(A) - \sum(\forall the child sets of B) \quad (4)$$

Finally, if we have k classes, N instances in the data set, the entropy of the entire set is E, the entropy of each of the subsets is E1 and E2, the number of instances in one class is *k*1 and in the other *k*2, ethen the minimum contribution of information is defined as (Ben & Bargaoui, 2020).

$$\frac{\log_2 N - 1}{N} + \frac{\log_2 3^k - 2 - K^E + k * E1 * E2}{N} \quad (5)$$

Algorithmk-means

K-means is one of the most widespread algorithms for grouping. Clustering represents a technique implemented in Data Mining. The idea of k-means is to place all objects in a given space and given their characteristics form groups of objects with similar but different features to the others that make up other groups. K-means is an unsupervised learning algorithm that has the following characteristics (Bai et al. 2017).

- The data set is partitioned into K groups (clusters) randomly.
- K center points are selected at random, one from each group (centroids).
- For each data, the distance from the point to each central point of the groups is calculated and the data becomes part of the group whose distance is less than its center.
- If the data is closer to its own group, it stays in its group, otherwise it becomes part of the group of the closest center.
- The previous process is repeated until no point passes the group.

However, the algorithm has some drawbacks:

- The final grouping depends on the initial centroids.
- Convergence in the global optimum is not guaranteed, and for problems with many specimens, it requires a large number of iterations to converge.

Materials and methods

This section describes the implementation scenario based on the stored data set of the tourism sector in the City of Riobamba.

Table 1. Data stored from the tourist sector in the City of Riobamba.

Data set	Type of data	Instances	Attributes
Valuation of fate	Multivariate	1382	4
Historical Tours	Multivariate	720	6
Hotel accommodation	Univariate	1080	5
Transport job	Multivariate	2801	4

A data set with decision attributes was used to be able to execute the algorithms excluding this attribute and then compare the results obtained with those originally indicated by these attributes.

By having data sets for which the decision attribute is known, it allows determining the number of groups of the k-means algorithm.

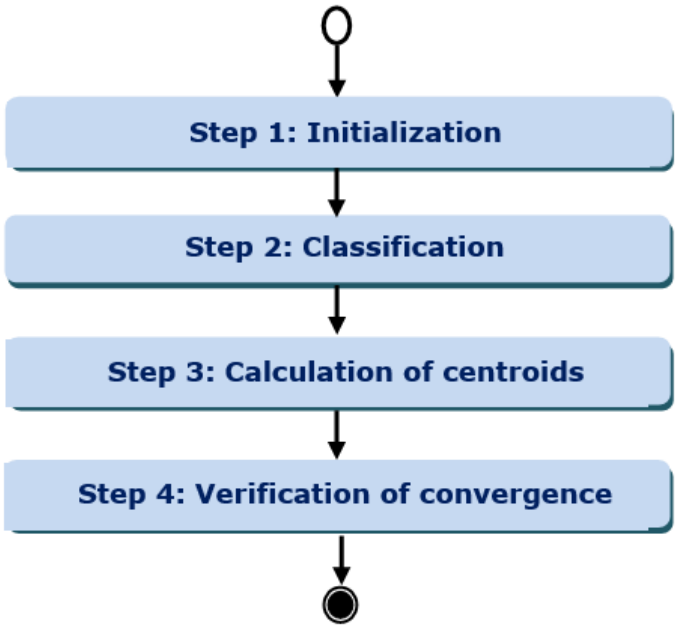

Figure 1 shows a diagram of the flow of the k-means method for the case under study.

Fig. 1. Diagram of the operation of the k-means method.

As in most data mining processes, each repository had to go through the stages of: clean, integrate, select, transform, mine, interpret and present (Mar et al. 2021). Figure 1 shows the flow of the k-means method for the case under study. The following is a description of the different steps that describe the method:

Step 1. Initialization: A set of objects is defined to which the clustering process is applied, which consists of dividing the data into groups and a centroid (geometric center of the clusters) for each one. Initial centroids can be determined randomly, while in other cases they process the data and centroids are determined by calculations.

Step 2. Classification: For each data, the square Euclidean distance from the centroids is calculated, the closest centroid to each of the data is determined, and the object is appended to the cluster of the centroid that was selected.

Step 3. Centroid calculation: The centroids are recalculated for each of the clusters.

Step 4. Convergence verification: It is checked if one of the algorithm's conditions has been met and that it must stop, this is called the convergence or stop condition. A set of conditions are defined for processing:

a) The number of iterations.

b) When the centroids obtained in two successive iterations do not change their value.

c) When the difference between the centroids of two successive iterations does not exceed a certain threshold.

d) When there is no transfer of objects between groups in two successive iterations.

If any of the convergence conditions is not met, steps two, three and four of the algorithms are repeated.

For computational processing the algorithms were coded in Python 3.8.12 and it was run on the following platform:

- Intel(r) core(tm) i3-2100 cpu @ 3.10ghzprocessor.
- Operating system: Ubuntu / Linux.

From the k-means algorithm comparison criteria, the one was chosen to maximize the number of success cases, since in the end the last interest is to determine how well the grouping did.

In order to compare the results obtained, three processes were run with the same data sets under the following conditions:

- 90 repetitions were made when random processes were carried out, in order to determine the average effect of the algorithm.
- When ranges were used, there was no point in repeating it more than once as the algorithm is deterministic for a given data set.

Classic k-means with random centroids

K-means was used as a grouping algorithm so that the resulting groups were then used to label the objects in their decision attribute (D); using the group number in which the object was grouped as the value of the decision attribute (D).

k-means using only the attributes with a contribution of information superior to a border. The entropy of each of the attributes and its information gain were calculated. The method used was as follows:

- Let E (C) be the entropy of the entire set of attributes.
- How much information is provided by the entropy of each of the ***c*** condition attributes (C) is calculated.
- Let ***E***(***ci***) be the entropy of the condition attribute ***ci***.
- As the selection of the criterion in which value, of the ***Vc*** values, to divide the attribute c to calculate the entropy can be very different for each attribute, it is decided to order the ***Vc*** values from least to greatest and take the mean as the division criterion.
- The information input of attribute ***c*** is equal to:
- The condition attributes that provide the greatest amount of information such as those selected are used

to choose the initial centers for the ***k-means*** algorithm from them.

Once the attributes to be considered have been chosen, if it is desired that the decision attribute (D) take different values from you, then ***k-means*** is run to form groups, using distances only the attributes selected for their greatest contribution of information. You can either initialize the centers randomly or divide the total range of the values of attribute ***c*** into ***k*** uniform pieces and take these values as initial centers of the ***k-means*** algorithm.

Given that they are the attributes that provide the most information, it was decided to initialize the centers with uniform ranges.

k-means using only attributes selected by rough sets

Using the theory of roughsets, to determine which condition attributes are indispensable and which are dispensable and therefore, proceed to the reduction of attributes, calculating the relation of indispensableness of each one of them.

Being P the set of attributes, a $\in P$, the attribute a is dispensable in P if:

$$IND(P)=IND(p\{a\}) \quad (6)$$

Similarly, once the attributes to be considered have been chosen, if it is desired that the decision attribute (D) take Vd different values, then k-means is run to form Vd groups using only the indispensable attributes for the calculation of distances. The centers can be initialized randomly, or the total range can be divided into k uniform pieces; In order to compare the results, the centers with uniform ranges were initialized.

RESULTS

From the data recovered from the tourist sector in the City of Riobamba, its processing is carried out. From the application of the previously proposed experiments, the results that are expressed in Table 2 are obtained.

Table 2. Results obtained for the different data sets.

Characteristic/ Data sets	Valuation of fate	Historical Tours	Hotel accommodation	Transport job
Total records	1382	720	1080	2001
Total attributes including decision	4	6	5	4
k-means classic				
Classic k-means success rate mean	63.40	35.20	53.20	54.50
Standard deviation of the classic k-means success rate	3.80	7.38	3.01	8.45
Variation coefficient of the classic k-means success rate	0.08	0.43	0.07	0.15
k-means using information gain				
Number of attributes removed due to information gain	1	1	1	1
Average success rate using the remaining attributes	56.80	30.40	56.20	48.30
Standard deviation of success rate using only the remaining attributes	0.00	0.00	0.00	0.00
Variation coefficient of success rate using only the remaining attributes	0.00	0.00	0.00	0.00
Rough sets				
Number of attributes removed by rough sets	2	1	0.00	0.00
Average success rate using the remaining attributes	57.6	32.32	54.20	42.40
Standard deviation of success rate using only the remaining attributes	0.00	0.00	0.00	0.00
Variation coefficient of success rate using only the remaining attributes	0.00	0.00	0.00	0.00

From the analysis of the results presented in Table 2, the following discussions are presented:

1. The classical ***k-means*** algorithm is highly dependent on the selection of the initial centers. Random center

initialization tends to have high standard deviations, therefore high coefficients of variation.

2. From the use of entropy and information gain, only the attributes that provide more information are used, uniform ranges are used for centroids instead of random centers. The process becomes deterministic for the same data set; therefore, the standard deviation and coefficient of variation are displayed at zero.
3. Once the data has been labeled, or if already labeled data sets are available, and although the determination of the indispensable and dispensable attributes using rough sets is an expensive process in computational time, once determined, the reduction of attributes benefits likewise the future classification process.
4. The classic ***k-means*** with random centers showed that in some cases it obtained a higher success rate than the others, the problem is that its standard deviation is high and, therefore, as the average case will not always occur, it can be perfectly give the worst case, or cases close to it, and in these scenarios its performance is lower than when using information gain or rough sets.

CONCLUSIONS

The present investigation proposed a machine learning method for dealing with unlabelled data sets those bases its operation on:

Use entropy and information gain to select from which attributes to calculate the k-means centers.

Use k-means with only the attributes selected from the previous step to label the data in your decision attribute.

Once the objects have been labeled with the previous steps, approximate sets can be used to determine which attributes are dispensable and which are indispensable and, therefore, proceed to the reduction of attributes.

The calculation of entropy, the information gain and the approximate sets requires a computational effort before calculating the k-means.

By implementing the k-means algorithm on the stored data set of the tourism sector in the City of Riobamba, a classification of the information is obtained from relevant data. The proposal provides a tool for decision-making based on obtaining better opportunities in the sector.

REFERENCES

Arnaiz, N. V. Q., Arias, N. G., & Muñoz, L. (2020). ***Neutrosophic K-means Based Method for Handling Unlabeled Data*** (Vol. 37). Infinite Study.

Baalaji, K., & Khanaa, V. (2020). A Review on Process of Data Mining Approaches in Healthcare Sectors. ***Executive Editor***, ***11***(01), 80.

Bai, L., Cheng, X., Liang, J., Shen, H., & Guo, Y. (2017). Fast density clustering strategies based on the k-means algorithm. Pattern Recognition, 71, 375-386.

Ben Jaafar, A., & Bargaoui, Z. (2020). Generalized Split-Sample Test Interpretation Using Rainfall Runoff Information Gain. Journal of Hydrologic Engineering, 25(1), 04019057.

Brachtl, M. V., Durant, J. L., Perez, C. P., Oviedo, J., Sempertegui, F., Naumova, E. N., & Griffiths, J. K. (2009). Spatial and temporal variations and mobile source emissions of polycyclic aromatic hydrocarbons in Quito, Ecuador. ***Environmental Pollution***, ***157***(2), 528-536.

Carrillo, H., Dames, P., Kumar, V., & Castellanos, J. A. (2018). Autonomous robotic exploration using a utility function based on Rényi's general theory of entropy. Autonomous Robots, 42(2), 235-256.

Crevecoeur, G. U. (2019). Entropy growth and information gain in operating organized systems. AIP Advances, 9(12), 125041.

Fraser, J. M., & Yu, H. (2021). Approximate arithmetic structure in large sets of integers. Real Analysis Exchange, 46(1), 163-174.

García, M. K. C., Carrillo, J. A. R., Francisco, J. M. P., Grunauer, M. S. N., & Berrezueta, L. A. O. (2017). Rural tourism and its sociocultural impact, by Casacay Parish residents, el 0ro Province, Ecuador. ***Revista Interamericana de Ambiente y Turismo-RIAT***, ***13***(2), 157-163.

Isler, S., Sabzevari, R., Delmerico, J., & Scaramuzza, D. (2016). An information gain formulation for active volumetric 3D reconstruction. In ***2016 IEEE International Conference on Robotics and Automation (ICRA)*** (pp. 3477-3484). IEEE.

Machado, S. (2018). Approximate lattices and Meyer sets in nilpotent Lie groups. ***arXiv preprint arXiv:1810.10870***.

Mar Cornelio, O., Gulín González, J., Bron Fonseca, B., & Garcés Espinosa, J. V. (2021). Sistema de apoyo al diagnóstico médico de COVID-19 mediante mapa cognitivo difuso. ***Revista Cubana de Salud Pública***, ***46***, e2459.

Ricardo, J. E., Poma, M. E. L., Argüello, A. M., Pazmiño, A. D. A. N., Estévez, L. M., & Batista, N. (2019). Neutrosophic model to determine the degree of comprehension of higher education students in Ecuador. Neutrosophic Sets and Systems, 26, 54-61.

Sadri, A., Ren, Y., & Salim, F. D. (2017). Information gain-based metric for recognizing transitions in human activities. Pervasive and Mobile Computing, 38, 92-109.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* (pp. 3-18). IEEE.

Sierra, J. C. (2016). Estimating road transport fuel consumption in Ecuador. Energy Policy, 92, 359-368.