Article **Journal of Business Development Strategies**
June 2019 Vol.5 No.15 9-13

Proposal of a simple recommendation system for small and medium enterprises for decision making based on unsupervised learning

Propuesta de un sistema simple de recomendación para pequeñas y medianas empresas para la toma de decisiones basado en el aprendizaje no supervisado

URUETA-HINOJOSA, Daniel E.†*, LARA-VELÁZQUEZ, Pedro, GUTIÉRREZ-ANDRADE, Miguel A., DE LOS COBOS-SILVA, Sergio G.

Universidad Autónoma Metropolitana - Iztapalapa. Departamento de Ingeniería Eléctrica, Av. San Rafael Atlixco 186, Col. Vicentina, Del. Iztapalapa, Ciudad de México, C.P. 09340

ID 1er Autor: *Daniel E., Urueta-Hinojosa* / **ORC ID**: 0000-0002-8741-6978

ID 1st Coauthor: *Pedro, Lara-Velázquez* / **ORC ID:** /0000-0002-8596-8234

ID 2nd Coauthor: *Miguel A., Gutiérrez-Andrade* / **ORC ID:** 0000-0002-8633-8592

ID 3rd Coauthor: *Sergio G., De los Cobos-Silva* / **ORC ID:** 0000-0003-1262-6310

DOI: 10.35429/JBDS.2019.15.5.9.13 Received March 11, 2019; Accepted May 23, 2019

Abstract

Recommendation systems are generally complicated, due they search to increase their reach and robustness, they combine different artificial intelligence approaches mainly of supervised learning. A disadvantage of this type of systems is that they must have a prior classification to be able to train a system and after they can be able to make decisions in a simmilar way that a human would do it; however, the task of classification is often expensive because is needed to consult with experts the possible classification (also known as label) that should be given to a specific data; although this method can be profitable for large companies, it is not for small and medium companies. This is the reason which the present work shows a proposal of a simple system that does not need to have a previous classification, allowing it to be profitable for small and medium enterprises in decision making.

Recommendation system, Unsupervised learning, Economy

Resumen

Los sistemas de recomendación generalmente son complicados debido a que ellos buscan aumentar su alcance y robustez, ellos combinan diversos enfoques de la inteligencia artificial principalmente del aprendizaje supervisado. Una desventaja de este tipo de sistemas es que deben de tener una clasificación previa para poder entrenar a un sistema y así poder tomar decisiones como un humano; sin embargo, la tarea de clasificación suele ser costosa debido a que se debe de consultar con expertos la posible clasificación (también conocida como etiqueta) que se le debe de dar a un dato específico lo cual puede ser redituable para las grandes empresas, pero no para las pequeñas y medianas. Es por esta razón que el presente trabajo muestra una propuesta de un sistema simple que no necesite tener una clasificación previa, permitiendo así que el mismo pueda ser redituable para pequeñas y medianas empresas en la toma de decisiones.

Sistema de recomendación, Aprendizaje no survervisado, Economía

Citation: URUETA-HINOJOSA, Daniel E., LARA-VELÁZQUEZ, Pedro, GUTIÉRREZ-ANDRADE, Miguel A., DE LOS COBOS-SILVA, Sergio G. Proposal of a simple recommendation system for small and medium enterprises for decision making based on unsupervised learning. Journal of Business Development Strategies. 2019, 5-15: 9-13.

* Correspondence to Author (email: deurueta@xanum.uam.mx)
† Researcher contributing first author

Introduction

In recent years, artificial intelligence has increased its popularity and applications, a particular case is one that is applied to analyze data and from them create recommendation systems, a very particular case is the Netflix award, which in 2010 offered one million dollars to the best recommendation system for its platform.

Recommendation systems can be as complex as required because they can use different artificial intelligence approaches according to the requirement; The traditional approach to create a recommendation system is based on supervised learning in which a system is trained with a large amount of data that, in turn, has a large number of characteristics; what you get is a system that can take desitions as an expert would do it.

Although this approach is practical for cases in which there are labels for the data such as: good, regular, bad or expensive, intermediate, cheap etc. It is not so convenient to implement when you have a large amount of data without labeling.

Consider the example of a vineyard in which a wine with certain chemical characteristics is labeled based on the opinion of tasters as: good, regular or bad; we have that although for a renowned vineyard it may be profitable to hire tasters to analyze 20% of their wines and then taking into account the results they can train a recommendation system; we cannot apply this for a small or medium vineyard.

The objective of this paper is to show the developed proposal of a simple recommendation system for small and medium enterprises to make decisions based on unsupervised learning approach.

This is intended to enable small and medium enterprises to create their own model and thus they can reduce their costs with respect to the classification of their products.

In the present article is described the fundamentals, then the methodology and finally the results of the proposal are analyzed and discussed.

1. Clustering

Clustering algorithms or clustering methods are dividing a data set into groups such that members of the same group are more similar among them than others (Ripley, 2007). The number of groups can be predetermined or can be decided by the algorithm. Interpretation of results that is obtained according to experts who analyze and interpret the partition of the data.

1.1. *k*-means algorithm

The *k*-means clustering (MacQueen, 1967) is the most commonly used unsupervised machine learning algorithm, the goal of this algorithm is to set into *k* groups a given dataset, in this algorithm each cluster is represented by the center or means of the data points belonging to the cluster. The basic pseudocode is (Bilmes, 1998):

1. Begin
2. Randomly choose *k* cluster centers
3. While points stop changing assignment to centroids

Assign each data point to the nearest cluster center

Set the cluster centroids based on the average (mean) position of each centroid´s points

4. End While
5. End

An inherent difficulty of the present algorithm is that the value of *k* must be known in advance, this number represents the number of groups (classes), for some instances this number is already set. However, assuming that this number is not available, there are statistical techniques which allows that it can be calculated, one of the simplest method to implement is known as elbow's method

1.1.1. Elbow method for K-Means algorithm

The elbow method consists in the running of the *k*-means algorithm for a given instance in a range of defined values of *k*, for example, for *k* = 1 to 15 and for each result obtained from *k*, the sum of the square errors is calculated (SSE).

ISSN 2444-4960

SSE is the sum of the squared differences between each observation and its group's mean (Kassambara, 2017). It can be used as a measure of variation within a cluster. If all cases within a cluster are identical the SSE would then be equal to 0. The SSE is given by the equation:

$$SSE = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad (1)$$

Finally, the results obtained are plotting; if the line chart looks like an arm, then the "elbow" on the arm is the value of *k* that is the best.

2. Instances

2.1. Wine instance

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars (Waterhouse, 1998). The analysis determined the quantities of 13 constituents found in each of the three types of wines: the good wines, the regular wines and the bad wines. The attributes are:

- Alcohol
- Malic acid
- Ash
- Alcalinity of ash
- Magnesium
- Total phenols
- Flavanoids
- Nonflavanoid phenols
- Proanthocyanins
- Color intensity
- Hue
- OD280/OD315 of diluted wines
- Proline

2.2. Automobile dataset

This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is riskier than its price indicates. Cars are initially assigned a risk factor symbol associated with its price (Kibler, 1989). The attributes are:

- Symboling: -3, -2, -1, 0, 1, 2, 3.
- Normalized-losses: continuous from 65 to 256.
- Make
- Fuel-type: diesel, gas.
- Aspiration: std, turbo.
- Num-of-doors: four, two.
- Body-style: hardtop, wagon, sedan, Hatchback, convertible.
- Drive-wheels: 4wd, fwd, rwd.
- Engine-location: front, rear.
- Wheel-base: continuous from 86.6 120.9.
- Length: continuous from 141.1 to 208.1.
- Width: continuous from 60.3 to 72.3.
- Height: continuous from 47.8 to 59.8.
- Curb-weight: continuous from 1488 to 4066.
- Engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
- Num-of-cylinders: eight, five, four, six, three, twelve, two.
- Engine-size: continuous from 61 to 326.
- Buel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
- Bore: continuous from 2.54 to 3.94.
- Stroke: continuous from 2.07 to 4.17.
- Compression-ratio: continuous from 7 to 23.
- Horsepower: continuous from 48 to 288.
- Peak-rpm: continuous from 4150 to 6600.
- City-mpg: continuous from 13 to 49.
- Highway-mpg: continuous from 16 to 54.
- Price: continuous from 5118 to 45400.

3. Evaluation of the model

3.1. Confusion matrix

To evaluate a model, it is frequently used the accuracy, defined as the ratio of correct predictions made by the model and the overall predictions. Given by the formula:

$$Accuracy = \frac{Total\ Correct\ Predictions}{Total\ Predictions} \quad (2)$$

Although this form is practical, it does not provide all the important information such as the total of correct and incorrect predictions made by the model.

ISSN 2444-4960

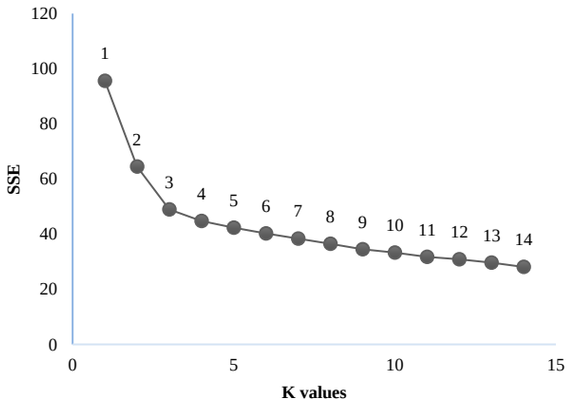
The confusion matrix was first known under the term Contingency Table; it was invented at 1904 by Karl Pearson 20. However, the term confusion matrix became popular in the machine learning community thanks to Kohavi and Provost at 1998. The confusion matrix generated from an instance with *n* classes, is a *nxn* square matrix where rows are named taking the real classes and columns, using the classes provided by the model. In this way, it can be clearly identified when the model classifies a class correctly or incorrectly. Thus, the confusion matrix allows us to evaluate the performance of the model with respect to an instance. For example, for an instance with two classes, the matrix would look like the Table 1:

	Negative (Model)	Positive (Model)
Negative (Real)	True Negative (TN)	False Positive (FP)
Positive (Real)	False Negative (FN)	True Positive (TP)

Table 1 Confusion matrix for 2 classes

4. Results

4.1. Wine results



Graphic 1 Elbow method in the wine instance

The results given by the elbow method show in the graphic 1 that the accurate number of classes are 3,

51	**0**	**0**
0	**65**	**0**
0	**0**	**62**

Figure 1 Confusion matrix for wine instance using the proposed model

Taking into account the confusion matrix shown in the Figure 1, it is possible to see that the model allows to visualize that the model forms three large groups intuitively, this means that all the elements belonging to a group are very similar each other, so it is not necessary to analyze all the elements but only one element of each group formed and, once it is classified as a bad, regular or good wine, it is possible to classify all the other elements in the same way without analyze them. For example, if we analize one element from the first group, we'll find that this group belongs to the good wines, in a similar way if is analize one element from the others groups, we'll find that they belong to the regular and bad wines respectively.

59	**0**	**0**
3	**65**	**3**
0	**0**	**48**

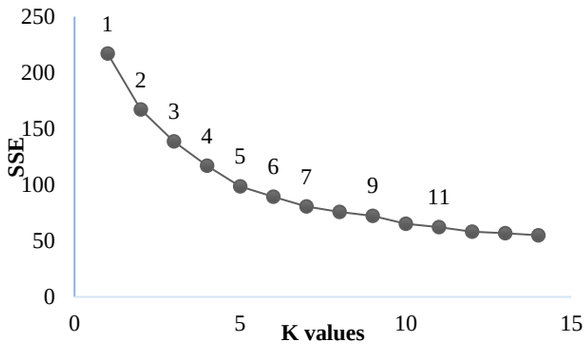
Figure 2 Confusion matrix for wine instance using the proposed model

The model also allows to evaluate a pre-classified instance, for example, in the case of wines; in Figure 2 the confusion matrix for the proposed model is shown; it is observed that the model has a disagreement in six elements between its results and the results obtained by the tasters (label). If we take into account the matrix, the results indicate that three wines that are being classified by experts as bad, for the proposed model are wines of regular quality, if it is translated into economic matters they would mean losses for the company because thay are selling a cheaper wine than it should be. On the other hand, there are also three wines that are actually regular but the company is classifying them as good quality wines, this situation can represent a risk too, because if it is proven that they sell wines with an inferior quality at a high price it will generate a bad image, which in turn could result in losses.

Despite all the above, the results do not mean that the tasters or the model made a mistake but simply that those wines have chemical characteristics which make a wine better or worse than it really is and maybe they should be tasted again.

ISSN 2444-4960

4.2. Automobiles results



Graphic 2 Elbow method in the automobile instance

The results given by the elbow method show in the graphic 2 that the accurate number of classes are 6, despite of in the dataset description indicates that there are 7 kind of cars, in the instance only appears 6 of them; that is the reason why the elbow method gets this value.

7	**0**	**0**	**0**	**0**	**0**
0	**49**	**0**	**0**	**0**	**0**
0	**0**	**24**	**0**	**0**	**0**
0	**0**	**0**	**42**	**0**	**0**
0	**0**	**0**	**0**	**28**	**0**
0	**0**	**0**	**0**	**0**	**9**

Figure 3 Confusion matrix for automobiles instance using the proposed model

In the same manner that in the wine instance, the confusion matrix shown in Figure 3, it is possible to conclude that the model separates the dates into the given number of groups forming six main groups, with this is admissible to determine that all the elements belonging to a group are very similar each other, so it is not necessary to analyze all the elements but only one element of each group formed and then we can assign the same layer to the rest of them.

Conclusions

The proposed model fulfills the function of being simple and at the same time being able to classify the test instances.

It is flexible because it can be applied in databases with a previous label and for those which there is no indication of how to classify. We can conclude that if certain characteristics are taken into account and analyzed with the model, they can save classification costs since once the result of the system is obtained, analyzing only one of the samples that make up a group, it is possible to determine the classification of all the other elements belonging to the same group. Likewise, in the case of instances where there is a previous classification, the model serves to verify these classifications and if required, an additional analysis is made because certain products may be overvalued or undervalued, which each situation to its way represents economic losses.

References

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute.

Kassambara, A. (2017). Practical guide to cluster analysis in R: unsupervised machine learning (Edition 1). Frankreich: STHDA.

MacQueen J. B. (1967). "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297.

Kibler, D., Aha, D.W., & Albert,M. (1989). Instance-based prediction of real-valued attributes. Computational Intelligence, Vol 5, 51--57.

Ripley, B. D. (2007). Pattern recognition and neural networks. Cambridge university press.
Waterhouse A. L., Ebeler. S. E. (1998). “Chemistry of wine flavor.” American Chemical Society, Oxford University Press.

ISSN 2444-4960