Knowledge-Based Systems 106 (2016) 231–241

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Automatic labelling of clusters of discrete and continuous data with supervised machine learning

Lucas A. Lopes[a], Vinicius P. Machado[a,*], Ricardo A.L. Rabêlo[a], Ricardo A.S. Fernandes[b], Bruno V.A. Lima[a]

[a] Department of Computer Science, Federal University of Piauí, Campus Universitário Ministro Petrônio Portella, Teresina 64049-550 PI, Brazil
[b] Department of Electrical Engineering, Federal University of São Carlos, Rod. Washington Luís (SP-310) km 235, São Carlos 13565-905 SP, Brazil

ARTICLE INFO

Article history:
Received 20 August 2015
Revised 21 May 2016
Accepted 23 May 2016
Available online 27 May 2016

Keywords:
Machine learning
clustering
labelling
artificial neural networks

ABSTRACT

The clustering problem has been considered one of the most relevant problems in the research area of unsupervised learning. However, the comprehension and definition of such clusters is not a trivial task, making necessary their identification, i.e., assign a label to each cluster. To address the problem of labelling clusters, this paper presents a methodology based on techniques for supervised learning, unsupervised learning and a discretization model. Thus, a method with unsupervised learning is applied to the clustering problem, and the supervised learning algorithm is responsible for detecting the meaningful attributes to define each formed cluster. Some strategies are used to form a methodology that presents a label (based on attributes and values) for each provided cluster. Such methodology is applied to three different databases, in which acceptable results were achieved with an average that exceeds 92.89% of correctly labelled elements.

1. Introdution

The clustering problem has been considered to be one of the most relevant problems in the research area of unsupervised learning, which is a subarea of Machine Learning (ML). Clustering refers to the process of partitioning a set of data (or objects) in smaller subsets, which are referred to as clusters or groups. The objects that possess similarities in their characteristics tend to belong to the same cluster, whereas objects with different characteristics tend to belong to different clusters.

The clustering process has been extensively used in several applications, such as business intelligence, image pattern recognition, web searches, biology, and security [1]. This way, in the business context, clustering can be used to organize a large number of customers into groups. This organization facilitates the development of business strategies for a better management of the relationship with the customer [2]. In a similar way, for web searches, clustering can be used to concisely organize information into similar subjects [3]. On the other hand, in the area of pattern recognition of images, clustering tools can be used to increase the accuracy of handwriting recognition systems [4].

According to [1], the clustering problem has been extensively explored irrespective of the area of application. However, the developed algorithms for this task exhibit the following problems:

- scalability - in many areas, the databases possess a large quantity of objects, such as millions or even billions of objects;
- knowledge representation - the ability to handle different types of attributes;
- nonlinearity - although some algorithms are limited to generating clusters of convex shape, many problems possess objects that are not linearly separable and require clusters with different shapes;
- problem domain - some techniques demand input parameters for their operation; for example, the number of clusters to be generated;
- noise - in many applications, the databases may contain objects with values that are unknown, non-existent or with error;
- new objects - some algorithms are capable of reorganizing their clusters according as new objects are presented, whereas other algorithms need to recalculate the entire process;
- high-dimensional data - some problems describe their objects with a large number of characteristics and the clustering techniques should be prepared to deal with this condition;
- constraints - in some cases, certain constraint conditions should be obeyed;

* Corresponding author.
E-mail address: vpmachado@gmail.com, vinicius@ufpi.edu.br (V.P. Machado).

http://dx.doi.org/10.1016/j.knosys.2016.05.044
0950-7051/

- interpretation and usability - experts require interpretable, understandable and useful results. The results are frequently related to the experts interpretation and the important characteristics during the process are not identified.

Although many researchers have focused on the development and improvement of algorithms for solving these problems, few studies are related to the interpretation of formed clusters [5–8]. According to [9], many researchers have been concerned with other problems and have not given the necessary attention to the specific problem of understanding the formed clusters.

This understanding is primarily due to the values presented by the most important characteristics of their objects. Thus, this set of relevant values represents a definition for any given cluster, that is, a label capable of providing experts with a better understanding of the problem.

For this reason, the interpretation of a label can imply several solutions or an optimization of the problem. As an example, we can cite the context of business intelligence, where an expert in a given company can easily identify the main characteristics that define different groups of customers from observation of the provided labels. Once the characteristics are known, the company can elaborate business models or specific strategies for a given group. Another example is a university that wishes to identify different groups of students and apply corrective actions that may increase student performance. The clusters can be easily formed, however, each cluster needs to be characterised. Thus, the existence of a label enables the identification of the characteristics that define a group.

Based on these above-mentioned examples, it is noticed that the understanding of clusters using labels can contribute to the development or optimization of the solution to a problem. Following this context, the objective of this paper is to determine a methodology for labelling clusters and, consequently, guide experts. The generated labels should identify the main attributes and the sets of values that are responsible for the definition of a given cluster.

Techniques for data clustering that involves distance metrics – for instance k-means – merely define the shortest distance between some input data and a centroid. Despite of, that in some cases, this methodology has proven being effective, the correlation criteria involving features and class are not clear making the process of labelling generated clusters an expensive and complex task. Furthermore, this would be an exhaustive task for a specialist since would be necessary a rigorous analysis for each generated cluster which may include a large number of attributes and ranges of values increasing drastically the approach complexity. Also, the relationship on each cluster is not always simple (linear) and may involve many features.

Hence, this work proposes a methodology able to deal and to present a better comprehension about clusters. The main aim here is to understand which combination of attributes and range of values mainly represents a cluster. Thus, the methodology proposed makes possible to determine some relationship between features on each cluster – that is, a label – without a thorough analysis from a specialist superimposing difficulties that emerges while trying to discover which combination of features and their values are most representative for clusters.

2. Theoretical framework

2.1. Machine learning

According to [10], the area of ML addresses the study of computational methods, which allow computer programs to autonomously improve in a given task by the use of experiments. In contrast to traditional computational methodologies, ML addresses the problem in such a manner that the machine will obtain after a learning process a hypothesis that defines the problem.

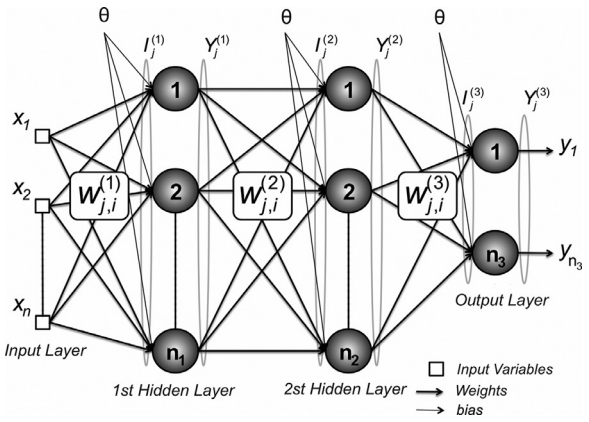

Fig. 1. A multilayer perceptron model (adapted from [11]).

Considering the area of ML from an extensive point of view, ML can be explained by two learning paradigms: supervised learning and unsupervised learning. In supervised learning, the creation of an accurate model is sought with regard to the prediction of values for new data. In unsupervised learning, the objective is to find a better approach to represent the data. Both cases involve a search for a model that is capable of generalizing unknown data and is differentiated by the existence of a label (response) that is present in the data used in supervised learning.

Due to the proposed methodology is based on Artificial Neural Networks and K-means algorithm, in the sequence, it will be given a brief introduction for such methods.

2.1.1. Supervised learning using artificial neural networks

In supervised learning, for each set of input values, a respective set of outputs exists that should be presented to the process. This process can be modelled as a classification problem, in which each sample contains a set of characteristics as the input and a set of responses as the output that are used during the stages of training and validation.

Following this context, this paper employs Artificial Neural Networks of the Multilayer Perceptron (MLP) type. According to [11], they represent computational models that have the ability to learn by using examples. Also, they are fault tolerant and have the ability to generalize solutions.

The ability to extract the existing relationships among the (several) variables of the problem via a training method will be explored in this study with regard to the detection of attributes that are meaningful to the problem, as detailed in Section 4 and in other studies, such as [12,13].

The Multilayer Perceptron (MLP) network is one type of ANN, which are commonly characterised by using at least one hidden layer (located between the input layer and the output layer). Moreover, the MLP uses a topology of the feedforward type, that is illustrated in Fig. 1.

The input signals of a MLP network (x_1, x_2, ..., x_n) have weights associated with each neuron of the next layer, which are represented by $w_{j,i}^{(1)}$, where j represents the neuron and i represents the input. The input value I of each neuron j belonging to the first hidden layer is calculated by the sum of the products of the input values of the previous layer according to Eq. (1):

$$I_j^{(1)} = \sum_{i=0}^{n} x_i \cdot w_{j,i}^{(1)}. \quad (1)$$

Next, the input value of a neuron is applied to an activation function $g(.)$. The most used function is generally the hyperbolic