Information Sciences 301 (2015) 181–194

Contents lists available at ScienceDirect

Information Sciences

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

Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods

Daniel Gomes Ferrari *, Leandro Nunes de Castro

Natural Computing Laboratory (LCoN), Mackenzie University, Brazil

ARTICLE INFO

Article history:
Received 20 March 2014
Received in revised form 13 December 2014
Accepted 27 December 2014
Available online 3 January 2015

Keywords:
Clustering
Problem characterization
Algorithm ranking
Algorithm selection
Meta-knowledge
Meta-learning systems

ABSTRACT

Data clustering aims to segment a database into groups of objects based on the similarity among these objects. Due to its unsupervised nature, the search for a good-quality solution can become a complex process. There is currently a wide range of clustering algorithms, and selecting the best one for a given problem can be a slow and costly process. In 1976, Rice formulated the Algorithm Selection Problem (ASP), which postulates that the algorithm performance can be predicted based on the structural characteristics of the problem. Meta-learning brings the concept of learning about learning; that is, the meta-knowledge obtained from the algorithm learning process allows the improvement of the algorithm performance. Meta-learning has a major intersection with data mining in classification problems, in which it is normally used to recommend algorithms. The present paper proposes new ways to obtain meta-knowledge for clustering tasks. Specifically, two contributions are explored here: (1) a new approach to characterize clustering problems based on the similarity among objects; and (2) new methods to combine internal indices for ranking algorithms based on their performance on the problems. Experiments were conducted to evaluate the recommendation quality. The results show that the new meta-knowledge provides high-quality algorithm selection for clustering tasks.

1. Introduction

It is well known that no single algorithm achieves the best performance over all instances of a problem class [75,76,80]. Rice [70] formulated the *Algorithm Selection Problem* (ASP), which proposes that there is a relation between the characteristics of a problem and the performance of the algorithm that can be used to solve it. The ASP is considered an NP-hard problem and has been tackled in different research fields [20,64].

The *meta-learning* field deals with the ASP by learning about the behavior of the learning algorithms [2,52,57,74]. A meta-learning system aims to learn which problems' characteristics contribute to a better performance of one algorithm over others [32,69] and, from this knowledge (*meta-knowledge*), to select the most suitable algorithm for a new, unseen problem [12].

The meta-knowledge, also known as *meta-data*, can be composed of the *meta-attributes* and the *meta-target* [12]. The meta-attributes are the characteristics or features extracted from the problems. The meta-target is the target variable for the meta-learning system [15,52]. Concerning the ASP, the meta-target may be the performance to be estimated or the rank

* Corresponding author.
E-mail addresses: ferrari.dg@gmail.com (D.G. Ferrari), lnunes@mackenzie.br (L.N. de Castro).

http://dx.doi.org/10.1016/j.ins.2014.12.044
0020-0255/

to be recommended for an algorithm [12,78]. An extensive review of meta-learning and the ASP can be found in Smith-Miles [75,76], in which the author presented how their integration is applied to different research fields, such as time series prediction [5,68], sorting [34], and optimization [75,76], among others.

Clustering is an unsupervised data mining task in which the goal is to find groups of similar objects [1,46]; the objects in the same group are more similar to one another than to objects from other groups. There are currently countless clustering algorithms for data mining tools. However, there is a lack of guidelines in the selection of algorithms for the analysis of a new problem [12].

The connections between data mining and meta-learning have been widely investigated for supervised tasks, such as classification [57,71,75,76]; however, there is no study, for instance, about the best meta-attributes to be extracted from unsupervised learning problems [12]. New data mining fields are applying meta-learning techniques to improve the performance of new methods, such as ensembles [49,51], mining data stream [72], mining big data [54], and among others.

Despite the large number of works related to the algorithm selection problem in the meta-learning literature, there are still grand challenges to be overcome [10]. The costly problem characterization and algorithm evaluation processes, together with the high-dimensional data sets, demand new ways to characterize and evaluate problems [10].

The works that tackle the algorithm selection problem by using meta-learning systems in the clustering context have in common the use of external indices to evaluate the algorithms. In other words, the clustering problems have previous known solutions, and these are used to quantify the proposed solutions. This common ground among these works makes it harder to extend the meta-knowledge because real-world clustering tasks usually do not have a priori known solutions.

In de Souto et al. [22], the authors used 32 microarray data sets about cancer gene expression as problems and 7 clustering algorithms. The problems were characterized by 8 meta-attributes, including one specific to the technology used in the construction of the microarray, and the algorithms were evaluated by an external index because the solution was already known. With a regression technique used to select the algorithm, the work presented good results. In Nascimento et al. [61], the authors applied meta-learning techniques to construct rules to automatically select clustering algorithms for gene expression data. A collection of 35 data sets was characterized by 13 meta-attributes, some of which were related to the size of existing clusters. By using an external index, 7 algorithms were evaluated by varying the similarity measure. The experiments aimed to find the best algorithm for rule extraction.

Soares et al. [77] used 160 artificially generated data sets characterized by 9 meta-attributes; the experiments evaluated two algorithms for ranking prediction: a neural network and a regression technique. The ranking was built with 9 clustering algorithms by using an error rate because the object labels were known in advance. In Ferrari and de Castro [26,27], the authors built a meta-learning system with 30 problems and 10 meta-attributes; 5 clustering algorithms were evaluated by using an external measure. Three estimation techniques were tested to predict the performance of the algorithms, and four classifiers were evaluated in a ranking recommendation task.

By taking into account these few related works, the present paper carries out an investigation into the meta-knowledge for clustering tasks. It proposes a new unsupervised characterization scheme based on the distance of objects; that is, an approach to obtain the meta-attributes that does not take into account the labels of objects but, instead, their distance, thus maintaining the unsupervised nature of the clustering task. Besides, the algorithms are evaluated by using different internal indices, which do not need a known solution, and they are combined by means of two new ranking combination methods.

To validate the proposals, a collection of 84 problems, 7 algorithms, and 10 internal indices are used, and the meta-knowledge is built with three distinct sets of meta-attributes: the traditional approach, the new unsupervised distance-based method, and a combination of both characterizations. Three ranking combination techniques are used for performance assessment: an existing method that uses the average rank position and two new methods based on score and competition. Then, this knowledge is applied to a meta-learning system to learn the relation between the problems' features and the performance of the algorithms, and the meta-knowledge quality is assessed by a selection mechanism.

The paper is organized as follows. Section 2 presents a meta-learning system for algorithm selection with the classic approach to obtain the meta-knowledge. Section 3 presents the main contributions of this work to the meta-knowledge of clustering tasks. Section 4 describes the experiments and the results obtained by the meta-learning system for the algorithm selection problem. Section 5 provides a discussion about the results presented and avenues for future research.

2. Meta-learning system

Building a meta-learning system to deal with the ASP requires the meta-knowledge from the learning process. This involves the following steps: collecting problems, choosing algorithms and evaluation measures, extracting meta-attributes, and determining the performance of the algorithms [12,52]. Then, when a new problem is presented to the meta-learning system, the meta-attributes are obtained, and a selection mechanism that makes use of the meta-knowledge database provides the ranking of the algorithms for the unseen problem.

Obtaining the meta-knowledge is a crucial step for the success of a meta-learning system and has been the subject of study in different research fields in the machine-learning community [12,47,52,75,76]. In the present paper, the concept of meta-knowledge is used as meta-data; i.e., the meta-knowledge is stored as an object composed of meta-attributes, which characterize the problems, and the ranking, which indicates the performance of the algorithms.