Contents lists available at ScienceDirect

Information Sciences

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

Towards supporting expert evaluation of clustering results using a data mining process model

Kweku-Muata Osei-Bryson

Department of Information Systems and The Information Systems Research Institute, Virginia Commonwealth University, Richmond, VA 23284, USA

ARTICLE INFO

Article history:
Received 11 August 2008
Received in revised form 13 August 2009
Accepted 25 September 2009

Keywords:
Cluster quality
Clustering goals
Expert evaluation
Decision support
Similarity measures
Data mining process model
CRISP-DM

ABSTRACT

Clustering is a popular non-directed learning data mining technique for partitioning a dataset into a set of clusters (i.e. a segmentation). Although there are many clustering algorithms, none is superior on all datasets, and so it is never clear which algorithm and which parameter settings are the most appropriate for a given dataset. This suggests that an appropriate approach to clustering should involve the application of multiple clustering algorithms with different parameter settings and a non-taxing approach for comparing the various segmentations that would be generated by these algorithms. In this paper we are concerned with the situation where a domain expert has to evaluate several segmentations in order to determine the most appropriate segmentation (set of clusters) based on his/her specified objective(s). We illustrate how a data mining process model could be applied to address this problem.

1. Introduction

The application of data mining techniques is becoming increasingly important in modern organizations that seek to utilize the knowledge that is embedded in the mass organizational data to improve efficiency, effectiveness and competitiveness. In recent years data mining practitioners and researchers have become aware of the need for formal data mining process models that prescribes the journey from data to discovering knowledge. Thus a multi-industry collective of practitioners (e.g. www.crisp-dm.org; [55]) came together to develop the cross-industry standard procedure for data mining (CRISP-DM) which was further extended by researchers (e.g. [17]). The data mining process (see Fig. 1) has been described in various ways (e.g. [17,55]) but essentially consists of the following steps [40]: Application Domain or Business Understanding (which includes definition of data mining goals), Data Understanding, Data Preparation, Data Mining, Evaluation (e.g. evaluation of results based on DM goals), and Deployment.

In this paper we will use the term *segmentation* to refer to the set of clusters that result from a given partitioning of the dataset by a clustering algorithm. It should be noted that a common presumption that underlies the proposal of solution approaches for clustering is that for each dataset there is a single optimal *segmentation* (i.e. partitioning) that is independent of the objectives of the end-user. For example Halkidi et al. [32] speaks of the " '*optimal' clustering scheme as the outcome of running a clustering algorithm (i.e., a partitioning) that best fits the inherent partitions of the data set*". So how would this 'optimal' partitioning (i.e. segmentation) be identified if two or more segmentations appear to have approximately the 'best' fit. Further would this choice be based on the goals of the clustering exercise (e.g. fraud detection vs. marketing profiling vs. theory building)? It appears to us that when clustering is used for knowledge discovery that what is the 'best fit' may not be independent of the goals of the clustering exercise. Thus Kim [38] suggests that for some situations "*comparison between*

E-mail address: Kweku.Muata@isy.vcu.edu

0020-0255/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2009.09.019

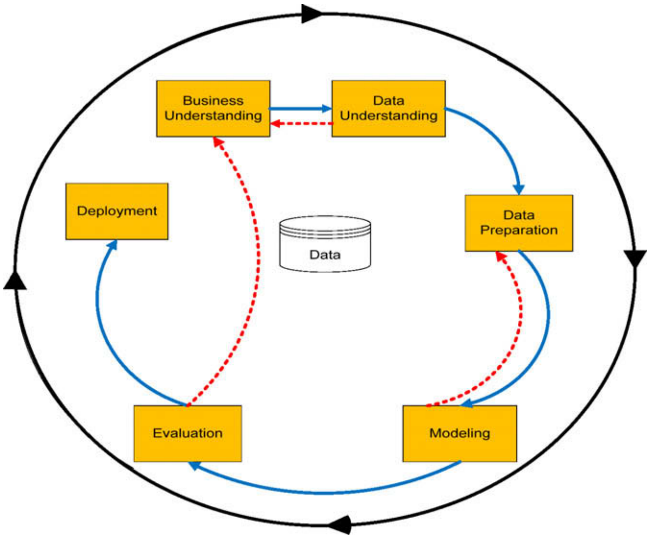

Fig. 1. CRISP process model (CRISP-DM, 2003).

clustering outputs of two methods should be interpreted only from managerial perspectives, rather than from numerical perspectives based on evaluation metrics such as intra-cluster compactness or inter-cluster separability". This position has been recognized in many data mining process methodologies including CRISP-DM ([55]; www.crisp-dm.org) where the first phase is Business Understanding (which includes the definition of Data Mining Goals and Success criteria), and later phases such as Modeling (i.e. Data Mining) and Evaluation being based on the Data Mining Goals and Success criteria. For as noted by Kurgan and Musilek [40], with regards to data mining "*The general research trends were concentrated on the development of new and improved DM algorithms rather than on the support for other KD activities*...*Before any attempt can be made to perform the extraction of this useful knowledge, an overall approach that describes how to extract knowledge needs to be established*". This statement is relevant for data mining in general, and in particular for undirected learning techniques such as clustering, where often it is assumed that there is in fact a single partitioning (i.e. segmentation) that is the appropriate one for a given dataset irrespective of subjective factors. For example, Halkidi et al. [32], listed the following steps for the clustering process: Variable Selection, Clustering Algorithm Selection, Validation of Results, and Interpretation. The reader may observe that although a process was specified for the clustering exercise that it did not include the specification of user-specified goals for the clustering exercise. On the other hand more recent DM process methodologies (e.g. [17,40,55]) assume that for clustering and other data mining problems, particularly those that occur within the business context, that there are 'subjective', problem specific clustering goals and success criteria beyond measures that are related to internal assessment of cluster quality.

In this paper we focus on illustrating how a generic data mining process model could be adapted for clustering. While in an earlier work [50] we focused on the case where only decision trees were used to describe the segmentation (i.e. set of clusters), in this work we focus on the case where assessment is done explicitly by the domain expert(s) in order to determine the most appropriate segmentation based on the specified goals of the data mining project. The fundamental assumption is that while the domain expert possesses the skill to make a good evaluation as to the appropriateness of a given segmentation, he/she is unable to evaluate more than a specified number of segmentations despite the fact that a large number of them might have been generated. The aim is to develop a technique that could empower domain experts to do this assessment in a non-burdensome way. This would offer a positive contribution to the DM process, and would encourage these domain experts to do thorough experimentation and analysis without being overwhelmed by the task of analyzing a significant number of segmentations. Given that this aspect of the DM process has not been adequately addressed by the DM research community (e.g. [2,37,32]), the results of this project could be valuable to DM researchers and practitioners.

Zopounidis and Doumpos [61] noted that "*When considering a discrete set of alternatives described by some criteria, there are four different kinds of analyses that can be performed in order to provide significant support to decision-makers: (1) to identify the best alternative or select a limited set of the best alternatives, (2) to construct a rank ordering of the alternatives from the best to the worst ones, (3) to classify/sort the alternatives into predefined homogenous groups, (4) to identify the major distinguishing features of the alternatives and perform their description based on these features*". Within this context, our research problem can be considered to involve selecting *a limited set of the 'best'* segmentations by the domain expert.

The adaptation of the generic data mining process model to clustering requires the availability of several procedures (e.g. internal assessment of cluster validity, automatic pruning of the set of segmentations based on the clustering goals). In addition to addressing our major research problem, other contributions of this paper include new procedures for: internal assessment of cluster validity; automatic pruning of the set of segmentations based on the three illustrative goals; assessment of