Unsupervised Ranking and Characterization of Differentiated Clusters

Luca Cazzanti, Courosh Mehanian, Julie Penzotti, Doug Scott, Oliver Downs
Contextual Marketing Team
Globys, Inc.
Seattle, WA, USA

***Abstract*— We describe a framework for automatically identifying and visualizing the most differentiating attributes of each cluster in a clustered data set. A dissimilarity function measures the cluster-conditional distinguishing saliency of each attribute with respect to a reference realization of the same attribute. For each cluster, the N attributes that are most dissimilar are presented first to the human expert, along with the overall dissimilarity of the cluster. We discuss the computational benefits of the proposed framework, how it can be implemented with map-reduce, its application to the behavioral analysis of mobile phone users, and it broad applicability to diverse problem domains.**

Keywords—clustering; dissimilarity; KL divergence; map-reduce;

I. Motivation

Given a data set divided into pre-defined clusters, human experts are often faced with the daunting tasks of identifying and describing the salient characteristics that most distinguish each cluster. “Big Data,” with its high dimensionality and large volumes, overwhelms the human analyst, who must resort to simplifying heuristics and subjective criteria that make the assessment of the differentiating saliency of each attribute more manageable. The subjectivity of these simplifying criteria may introduce biases and inconsistencies in the saliency assessment, and mislead the analysts to erroneous conclusions.

A related challenge is that the clustering is typically obtained from a small subset of the attributes that characterize each element in the data set, but the analyst must assess the cluster-conditional, distinguishing saliency of all available attributes, and thus faces a combinatorial explosion of possible solutions. This challenge is further compounded by the need to assess the distinguishing saliency of vector-valued attributes, which places an additional burden on the human expert’s cognitive processes.

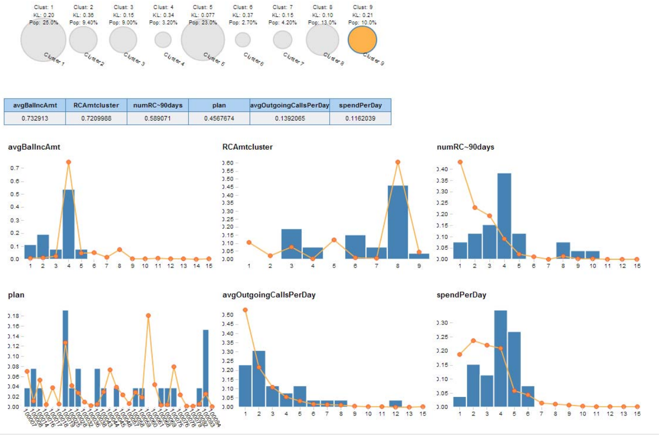
II. Proposed Method

Our proposed framework meets these challenges by automatically and objectively measuring how differentiated each cluster-conditional attribute is from a chosen reference attribute using mathematical measures of dissimilarity. The overall dissimilarity of each cluster can be computed by appropriately aggregating the individual attribute dissimilarities.

In the proposed framework, the chosen dissimilarity measure is the KL divergence between the cluster-conditional attribute and the same attribute for the entire data set. The overall cluster dissimilarity is computed by averaging the KL divergence values of all the attributes.

For a given clustering of *C* clusters and *A* different attributes, the proposed method requires *CxA* evaluations of the KL divergence, which can be parallelized trivially across multiple processors on modern, cloud-computing resources using the map-reduce computational paradigm. Similarly, ordering the cluster-conditional attributes by their KL divergence can be done in parallel, and the overall cluster dissimilarity is computed in the final reduce step of map-reduce, which averages the individual attribute KL divergences.

We apply the proposed framework to the behavioral analysis of prepaid mobile phone customers, clustered according to their phone usage patterns. However, the proposed framework flexibly accommodates any dissimilarity function and any aggregation strategy for computing the cluster dissimilarity, which makes it applicable to a wide variety of problem domains. Possible applications to security include identifying and most the unusual patterns from a large set of candidate attributes for a given clustering of human activity patterns.



avgBalIncAmt	RCAmtcluster	numRC~90days	plan	avgOutgoingCallsPerDay	spendPerDay
0.732913	0.7209988	0.589071	0.4567674	0.1392065	0.1162039

Fig. 1: Visualization of the 6 most differentiated attributes of a chosen cluster, with respect to the corresponding attributes of the overall population. The attributes are selected automatically.