Constrained cluster size for K-prototypes

anonymous

***Abstract*—In the real-world application, the data features are usually provided as the mixture of catogorical and numerical ones. The current K-prototypes can be applied to solve the clustering problem. However, the traditional K-prototypes has no control on the cluster size, which may lead to some clusters could contein very few nodes. To overcome this problem, we propse two algorithms based on the K-prototypes with cluster size constraints. The numerical results showed our proposed methods can cluster the nodes for a better summary of the unlabeled data.**

I. Introduction

In contemporary data analysis and machine learning applications, datasets often present a complex amalgamation of both categorical and numerical features. This heterogeneity poses a unique set of challenges when it comes to clustering and categorizing data points effectively. Clustering, as a fundamental technique in data analysis, plays a crucial role in uncovering hidden patterns, enhancing decision-making processes, and facilitating the organization of large datasets. Among the myriad of clustering techniques available, K-prototypes has emerged as a prominent choice due to its ability to handle mixed data types, i.e., both categorical and numerical.

The traditional K-prototypes algorithm is indeed a valuable tool, enabling the segmentation of data points into clusters based on their similarity. However, one critical limitation of the conventional K-prototypes approach lies in its inability to regulate cluster sizes. This limitation may result in some clusters comprising only a handful of data points, which, in practice, can lead to suboptimal clustering outcomes and less informative data summaries.

Recognizing the significance of addressing this shortcoming, we present in this paper two novel algorithms rooted in the K-prototypes framework, each designed to incorporate cluster size constraints. These innovative methods aim to enhance the clustering process by offering greater control over the sizes of the resulting clusters, ensuring that clusters contain a more balanced distribution of data points.

In this introductory section, we provide an overview of the challenges posed by mixed data types and the importance of clustering in data analysis. We also introduce the traditional K-prototypes algorithm and highlight its strengths and limitations. Subsequently, we outline the primary motivation behind our research – the need for cluster size control – and briefly preview the two proposed algorithms. Finally, we offer a glimpse into the numerical results of our experiments, which demonstrate the effectiveness of our methods in achieving improved clustering solutions for summarizing unlabeled data.

The organizaiton of this papaer is the following: in Sec.II, we discuss the problem statement. Then the Sec.III talks about the proposed algorithms and theretical analysis.The numerical experiments are shown in Sec.IV. The final section Sec.V presents our conclusions.