Abstract

In this thesis, we study the k-median problem with respect to a dissimilarity measure D_φ from the family of Bregman divergences: Given a finite set P of size n from $\mathbb{R}^d$, our goal is to find a set C of size k such that the sum of error $\mathrm{cost}(P,C) = \sum_{p\in P} \min_{c\in C}\{\mathrm{D}_\varphi(p,c)\}$ is minimized. This problem plays an important role in applications from many different areas of computer science, such as information theory, statistics, data mining, and speech processing.

Our main contribution is the development of a general framework of algorithms and techniques that is applicable to (almost) all Bregman divergences. In particular, we give a randomized approximation algorithm for the Bregman k-median problem that computes a $(1+\varepsilon)$-approximate solution using at most $2^{\tilde{\mathcal{O}}(k/\varepsilon)}n$ arithmetic operations, including evaluations of Bregman divergence D_φ. In doing so, we give the first approximation algorithm known for this problem that provides any provable approximation guarantee. We also give a fast, practical, randomized approximation algorithm that computes an $\mathcal{O}(\log k)$-approximate solution for arbitrary input instances, or even an $\mathcal{O}(1)$-approximate solution for certain, well separated input instances.

In addition to that, we study the use of coresets in the context of Bregman k-median clusterings. In a nutshell, a coreset is a small (weighted) set that features the same clustering behavior as the original input set. We show how classical coreset constructions for the Euclidean k-means problem can be adapted to a special subfamily of the Bregman divergences, namely the class of Mahalanobis distances. We also give a new, randomized coreset construction for the Mahalanobis k-median problem in low dimensional spaces that has several practical advantages. Furthermore, by introducing the notion of weak coresets, we give the first coreset construction applicable to (almost) all Bregman k-median clustering problems. Using these weak coresets, we are able to give the currently asymptotically fastest $(1+\varepsilon)$-approximation algorithm known for the Bregman k-median problem. This algorithm uses at most $\mathcal{O}(kn) + 2^{\tilde{\mathcal{O}}(k/\varepsilon)}\log^{k+2}(n)$ arithmetic operations, including evaluations of Bregman divergence D_φ.