Online Clustering with Experts

Anna Choromanska
Department of Electrical Engineering
Columbia University
aec2163@columbia.edu

Claire Monteleoni
Department of Computer Science
George Washington University
cmontel@gwu.edu

Abstract

Approximating the k-means clustering objective with an online learning algorithm is an open problem. We introduce a family of online clustering algorithms by extending algorithms for online supervised learning, with access to expert predictors, to the unsupervised learning setting. Instead of computing prediction errors in order to re-weight the experts, the algorithms compute an approximation to the current value of the k-means objective obtained by each expert. When the experts are batch clustering algorithms with b-approximation guarantees with respect to the k-means objective (for example, the k-means++ or k-means# algorithms), applied to a sliding window of the data stream, our algorithms obtain approximation guarantees with respect to the k-means objective. The form of these online clustering approximation guarantees is novel, and extends an evaluation framework proposed by Dasgupta as an analog to regret. Our algorithms track the best clustering algorithm on real and simulated data sets.

1 Introduction

As data sources continue to grow at an unprecedented rate, it is increasingly important that algorithms to analyze this data operate in the *online learning* setting. This setting is applicable to a variety of data stream problems including forecasting, real-time decision making, and resource-constrained learning. Data streams can take many forms, such as stock prices, weather measurements, and internet transactions, or any data set that is so large compared to computational resources, that algorithms must access it in a sequential manner. In the online learning model, only one pass is allowed, and the data stream is infinite.

Most data sources produce raw data (*e.g.* speech signal, or images on the web), that is not yet labeled for any classification task, which motivates the study of *unsupervised learning*. *Clustering* refers to a broad class of unsupervised learning tasks aimed at partitioning the data into clusters that are appropriate to the specific application. Clustering techniques are widely used in practice, in order to summarize large quantities of data (*e.g.* aggregating similar online news stories), however their outputs can be hard to evaluate. Probabilistic assumptions have often been employed to analyze clustering algorithms, for example i.i.d. data, or further, that the data is generated by a well-separated mixture of Gaussians.

Without any distributional assumptions on the data, one way to evaluate clustering algorithms, without having domain expert in the loop, is to formulate some objective function, and then to prove that the clustering algorithm either optimizes it, or is an approximation algorithm. Approximation guarantees, with respect to some reasonable objective, are therefore useful. The k-means objective is a simple, intuitive, and widely-cited clustering objective, however few algorithms provably approximate it, even in the batch setting. In this work, inspired by an open problem posed by Dasgupta [2], our goal is to approximate the k-means objective in the online setting.

1.1 The k-means clustering objective

One of the most widely-cited clustering objectives for data in Euclidean space is the k-means objective. For a finite set, S, of n points in $\mathbb{R}^d$, and a fixed positive integer, k, the k-means objective is to choose a set of k cluster centers, C in $\mathbb{R}^d$, to minimize:

$$\Phi_X(C) = \sum_{x \in S} \min_{c \in C} \|x - c\|^2$$

which we refer to as the "k-means cost" of C on X. This objective formalizes an intuitive measure of goodness for a clustering of points in Euclidean space. Optimizing the k-means objective is known to be NP-hard, even for $k = 2$ [1]. Therefore the goal is to design *approximation algorithms*. Surprisingly few algorithms have approximation guarantees with respect to k-means, even in the batch setting. Even the algorithm known as "k-means" does not have an approximation guarantee.

Our contribution is a family of online clustering algorithms, with regret bounds, and approximation guarantees with respect to the k-means objective, of a novel form for the online clustering setting. We extend algorithms from [3] and [4] to the unsupervised learning setting, and introduce a flexible framework in which our algorithms take a set of candidate clustering algorithms, as experts, and track the performance of the "best" expert, or best sequence of experts, for the data. Our approach lends itself to settings in which the user is unsure of which clustering algorithm to use for a given data stream, and exploits the performance advantages of any batch clustering algorithms used as experts. Our algorithms vary in their models of the time-varying nature of the data; we demonstrate encouraging performance on a variety of data sets.

2 Online k-means approximation

Our analysis is inpired, in part, by an evaluation framework proposed by Dasgupta as an analog to regret [2]. The *regret* framework, for the analysis of supervised online learning algorithms, evaluates algorithms with respect to their additional prediction loss relative to a hindsight-optimal comparator method. With the goal of analyzing online clustering algorithms, Dasgupta proposed bounding the difference between the cumulative clustering loss since the first observation:

$$L_T(\texttt{alg}) = \sum_{t \leq T} \min_{c \in C_t} \|x_t - c\|^2 \tag{1}$$

where the algorithm outputs a clustering, C_t, before observing the current point, x_t, and the optimal k-means cost on the points seen so far. We provide clustering variants of predictors with expert advice from [3] and [4], and analyze them by first bounding this quantity in terms of regret with respect to the cumulative clustering loss of the best (in hindsight) of a finite set of batch clustering algorithms. Then, adding assumptions that the batch clustering algorithms are b-approximate with respect to the k-means objective, we extend our regret bounds to obtain bounds on this quantity with respect to the optimal k-means cost on the points seen so far.

References

[1] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. Np-hardness of euclidean sum-of-squares clustering. *Mach. Learn.*, 75:245–248, May 2009.

[2] Sanjoy Dasgupta. Course notes, CSE 291: Topics in unsupervised learning. Lecture 6: Clustering in an online/streaming setting. Section 6.2.3. In *http://www-cse.ucsd.edu/∼dasgupta/291/lec6.pdf, University of California, San Diego, Spring Quarter*, 2008.

[3] M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.

[4] Claire Monteleoni and Tommi Jaakkola. Online learning of non-stationary sequences. In *NIPS*, 2003.

Topic: learning algorithms, learning theory Preference: oral/poster

Adresses:

Anna Choromanska
Computer Science
Columbia University
CEPSR 624, Mail Code 0401
1214 Amsterdam Avenue
New York, NY 10027

Claire Monteleoni
Department of Computer Science
The George Washington University
801 22nd St. NW, Suite 703
Washington DC, 20052