Imreh Csanád - Online klaszterezési algoritmusok

2012. okt. 05.

Online problems

The input is given part by part and the algorithm has to make the decisions without any information on the further parts.
The first published online problem is in the Greek mythology. The performance of an algorithm is measured by the competitive analysis or by an average case analysis.
An algorithm for a minimization problem is c-competitive if its cost is at most c - times more than the optimal cost.
The first analysis for an online scheduling algorithm was done by Graham in 1966. Since 1980 many results have been achieved and several areas have been developed.

Online unit covering

In unit covering, a set of n points needs to be covered by balls of unit radius, and the goal is to minimize the number of balls used.

Online unit clustering on line

In unit clustering the online algorithm is not required to fix the exact position of each ball in advance. The algorithm needs to make sure that a set of points which is assigned to one ball (cluster) can always be covered by that ball, thus the ball can be shifted if necessary.

- Chan and Zarrabi-Zadeh (2009) 2-competitive algorithm for line
- Chan and Zarrabi-Zadeh (2009) 16/11 –competitive randomized algorithm for line
- Epstein, van Stee (2010) 7/4-competitive algorithm for line, 8/5 lower bound on the possible competitive ratio for line
- Ehmsen, Larsen (2010) 5/3-competitive algorithm for line
- in two dimensional problems, usually the Linfinity norm is
- considered

Online facility location

In the facility location problem a metric space is given with a multiset of demand points (elements of the space). The goal is to find a set of facility locations in the metric space which minimizes the sum of the facility cost and assignment cost.

- Meyerson (2001): No constant competitive algorithm exists, An O(log n)-competitve randomized algorithm which is constant - competitive algorithm for randomly ordered inputs
- Fotakis (2003,2007): An O(log(n)/log log(n))-competitive algorithm and a matching lower bound on the possible competitive ratio.

- Anagnostopoulos et al (2004): A simpler O(log n)-competive algorithm, the first average case analysis
- Fotakis (2006) Divéki and Imreh (2010): Facility location with facility movements

Online clustering with variable sized clusters I.

The flexible model: In this model, when a new cluster is opened we need to specify its label, but its coordinates as well as its diameter might be changed by the algorithm in the future. For this model the cost of a cluster may change as new points are assigned to it.

The strict model: In this model, when a new cluster is opened we need to specify the coordinates of the interval which will be associated with this cluster, and the algorithm is allowed to assign only points belonging to this interval to the cluster. Here the cost of a cluster is defined as 1 plus the length of the interval associated with it.
The intermediate model: In this model, when a new cluster is opened we need to specify the length of the interval which will be associated with this cluster, but its coordinates might be changed by the algorithm in the future.
The algorithm cannot assign a new point to an existing cluster, if this will increase its diameter beyond the length which was specified for this cluster.

Algorithm: Extend Closest Clusters

Theorem: Algorithm Extend Closest Clusters fi-competitive.
Theorem: There is no deterministic online algorithm for the flexible model whose competitive ratio is strictly smaller than fi.

Increasing input sequence in the flexible model

Algorithm OnlOpt: When a new point arrives, and its distance from the last opened cluster is at least 1, the algorithm opens a new cluster and assigns the new point to the new cluster. Otherwise the point is assigned to the last opened cluster.
Theorem: Algorithm OnlOpt is 1-competitive.
Proof: To show that this algorithm results in an optimal solution, consider a fixed optimal solution OPT which maximizes the number of clusters (among all optimal solutions). One can show by induction on the number of points in a prefix of the input that the solution returned by the online algorithm is equal to OPT.

Increasing input sequence in the strict model

We consider the following simple semi-online algorithm. Upon arrival of a new request point p, if it is not already covered by a cluster, open the cluster [p; p + 1].
Theorem The competitive ratio of this semi-online algorithm is 2.
Proof: The cost of the algorithm is 2k where k is the number of used clusters. On the other hand there are k requests with pairwise distance at least 1, and this proves
that the optimal cost is at least k.

Theorem: The competitive ratio of any online algorithm on increasing input sequences for the strict model is at least 2.

2-dimensional versions

Given an input consisting of n requests which are points on the 2-dimensional plane, the goal is to partition the points into groups called clusters. We consider two cost functions.

In the case of linear cost, the cost of a cluster C is defined as $1 + \max_{i,j \in C} |i - j|_\infty$, that is, the sum of a fixed cost which is scaled to 1, and the l_∞ norm diameter of the cluster.

In the case of the square cost, the cost of cluster C is defined as $1 + \max_{i,j \in C} |i - j|_\infty^2$, that is, the sum of a fixed cost which is scaled to 1, and the square of the l_∞ diameter of the cluster.

String clustering

We have to divide a sequence of n-dimensional bitvectors into the minimal number of clusters where each cluster has diameter at most k by the Hamming distance.

In the online model the strings arrive one by one and after the arrival a string we have to assign it to an already existing cluster or to define a new cluster for it.

Greedy algorithm If the string can be assigned to a cluster assign to the first such cluster. Otherwise open a new cluster for it.

Theorem The competitive ratio of Greedy is $3/2$ if $k = 1$, and no online algorithm can have smaller competitive ratio than $3/2$.

Greedy for k=2

Theorem Greedy is $\Theta(n)$-competitive if $k = 2$.

Proof idea Since the optimal clusters contain at most $n + 1$ elements it is $O(n)$ -competitive.

But it is not better. Suppose that a sequence $(0, 0, x_i), (1, 1, x_i)$ arrives $i = 1, \ldots, n - 1$ where the set x_1, x_2, x_{n-1} is a set of $n - 2$ dimensional strings having diameter 2.

Then the greedy algorithm forms $n - 1$ clusters each containing two elements.

The optimal solution has two clusters one contains the strings started by $(0, 0)$ the other contains the strings started by $(1, 1)$.