Supplementary Material

Paper title: Explainable Clustering via Simultaneous Construction of Clusters and Exemplars: Complexity and Provably Good Approximation Algorithms

8 Additional Definitions

Graph Theoretic Definitions: We use some graph theoretic concepts and a special class of graphs in proving our results. Given an undirected graph $G(V, E)$, a subset V' of nodes forms a **dominating set** for G if for every node $w \in V - V'$, there is a node $v \in V'$ such that the edge $\{v, w\}$ is in E. Given a graph $G(V, E)$, the goal of the **minimum dominating set** (MDS) problem is to find a dominating set of minimum cardinality for G.

Given a set of disks (i.e., circles in two-dimensional space) each with the same radius r, one can define an associated undirected graph as follows: there is one node for each disk; there is an edge between two nodes if the corresponding disks touch or intersect (i.e., the distance between the centers of the disks is at most $2r$). Such a graph is called a **unit disk graph** [2]. Many optimization problems, including the MDS problem, are known to be **NP**-hard even for unit disk graphs [2,10]. We rely on the **NP**-hardness of the minimum dominating set problem for unit disk graphs in proving Proposition 4.1.

Unit disk graphs can also be defined in three or more dimensions. In such a case, each object is a ball of unit radius in an appropriate dimension. Each node of the corresponding graph represents a ball and there is an edge between two nodes if the corresponding balls touch or intersect.

Minimum Set Cover (MSC) Problem: In this problem [7], the input consists of a base set $U = \{u_1, u_2, \ldots, u_n\}$, a collection $Y = \{Y_1, Y_2, \ldots, Y_m\}$, where each Y_j is a subset of U $(1 \leq j \leq m)$ and an integer bound $\beta \leq m$. The goal is to choose a subcollection Y' of Y with $|Y'| \leq \beta$ such that the union of the sets in Y' is equal to U (i.e., the union covers all the elements in U). This problem is **NP**-complete and a natural greedy approximation algorithm (which picks a new set in each iteration such that the set covers as many new elements as possible) is known to give a performance guarantee of $O(\log n)$ for the problem [18]. One of our results (Section 4.3) uses this approximation algorithm.

Budgeted Maximum Coverage Problem: We also use a known approximation algorithm for the Budgeted Maximum Coverage (BMC) problem, which is closely related to the Minimum Set Cover (MSC) problem [7]. The input to the BMC problem is a base set $U = \{u_1, u_2, \ldots, u_n\}$, a collection $Y = \{Y_1, Y_2, \ldots, Y_m\}$, where each Y_j is a subset of U $(1 \leq j \leq m)$ and a budget $\beta \leq m$. The goal is to choose a subcollection Y' of Y with $|Y'| = \beta$ such that the union of the sets in Y' covers the maximum number of elements of U. This problem is also **NP**-hard and a natural greedy approximation algorithm (which picks a new set in each iteration such that the set covers as many new elements as possible) has been shown to give a performance guarantee of $(1 - 1/\mathrm{e})$ for the problem [12], with e being the base of the natural logarithm. One of our results (Section 4.3) uses this result.

9 Statement and Proof of Proposition 4.1

Statement of Proposition 4.1: The MSE problem is **NP**-hard even when the set of instances X consists of points in two-dimensional Euclidean space and the distance between any two points is their Euclidean distance.

Proof: The proof is by a reduction from the minimum dominating set (MDS) problem for unit disk graphs discussed in Section 8. Let the MDS problem be specified by a unit disk graph $G(V, E)$, where the radius of each disk is r, and let $\beta \leq |V|$ be the given upper bound on the size of a dominating set. We construct a set of instances X for the MSE problem as follows. For the disk corresponding to each vertex v_i, we create an instance $x_i \in X$, where the coordinates of x_i are those of the center of the disk corresponding to v_i. The exemplar distance ϵ is set to $2r$ and the bound on the number of exemplars is set to β. Obviously, this construction can be done in polynomial time.

Suppose V' is a dominating set for G with at most β nodes. We can show that the instances corresponding to the nodes in V' form the exemplar set $\mathcal{E}$ for X as follows. Consider any instance x_j in X which is not an exemplar. Since V' is a dominating set and the node v_j corresponding to x_j is not in V', there is a node $v_i \in V'$ such that the edge $\{v_i, v_j\}$ is in E. Since G is a unit disk graph, the distance between the centers of the disks corresponding to v_i and v_j is at most $2r$ which is equal to ϵ by our construction; that is, the distance between x_j and the exemplar x_i is at most ϵ. Therefore, $\mathcal{E}$ is a set of exemplars of size at most β for X.

Now, suppose $\mathcal{E}$ is a set of exemplars of size at most β for X. Let V' be the set of nodes of G corresponding to the instances in $\mathcal{E}$. We claim that V' is a dominating set for G. To see this, consider any node v_j which is not in V'. The instance x_j corresponding to v_j has an exemplar $x_i \in \mathcal{E}$ and the distance between x_i and x_j is at most $2r$. Since G is a unit disk graph, the edge $\{v_i, v_j\}$ is in E. In other words, V' is a dominating set for G, and this completes the proof. ■

10 Statement and Proof of Theorem 4.1

Statement of Theorem 4.1: The solution produced by Algorithm 1 satisfies the following conditions: (i) The diameter of each cluster is at most $2(D^* + \epsilon)$, where D^* is the optimal diameter for a k-clustering of X and ϵ is the exemplar distance. (ii) Every instance in X has an exemplar (at a distance of at most ϵ) within the same cluster. (iii) The sets of exemplars for the k clusters are pairwise disjoint. (iv) The total number of exemplars generated by the algorithm is at most $O(N^* \log n)$, where N^* is the minimum number of exemplars needed to cover all the instances in X.

Proof: To prove Part (i), we first note that the approximation algorithm used in Step 1 guarantees that the maximum diameter of the clusters produced in that step is at most $2D^*$, where D^* is the optimal solution value for X. Step 6 of the algorithm moves non-exemplars between clusters. We need to show that after these moves, the maximum diameter is at most $2(D^* + \epsilon)$. To see this, consider any cluster C_i and any pair of instances x_a and x_b in C_i. There are three cases to consider.

Case 1: Both x_a and x_b are exemplars. In this case, both x_a and x_b must be in B_i since we chose $\mathcal{E}_i = B_i \cap A$. Thus, at the end of Step 1, $d(x_a, x_b) \leq 2D^*$.

Case 2: One of them, say x_a, is an exemplar and the other (i.e., x_b) is a non-exemplar that got moved into C_i. In this case, C_i contains an exemplar x_q at a distance of at most ϵ from x_b. Since $d(x_a, x_q) \leq 2D^*$ and $d(x_q, x_b) \leq \epsilon$, it follows from triangle inequality that $d(x_a, x_b) \leq 2D^* + \epsilon$.

Case 3: Both x_a and x_b are non-exemplars which were moved into C_i. In this case, C_i contains exemplars x_p and x_q such that $d(x_a, x_p) \leq \epsilon$ and $d(x_b, x_q) \leq \epsilon$. Further, $d(x_p, x_q) \leq 2D^*$. Now, using triangle inequality, it follows that $d(x_a, x_b) \leq 2(D^* + \epsilon)$, and this completes our proof of Part (i).

The result in Part (ii) follows since the set A constructed in Step 3 is an exemplar set for X and each non-exemplar instance x_j gets moved (in Step 6) to a cluster containing an exemplar for x_j. Since the blocks constructed in Step 1 are pairwise disjoint, so are the exemplar sets constructed in Step 5; this proves Part (iii). Since Step 3 uses the greedy approximation algorithm for MSC and this algorithm provides a performance guarantee of $O(\log n)$ [18], the total number of exemplars produced in Step 3 is at most $O(N^* \log n)$, where N^* is the minimum number of exemplars needed to cover all the instances in X. This establishes Part (iv) and the theorem follows. ■

Expanded version of the Remark in Section 4.2: The remark in Section 4.2 mentions that one can theoretically get a better performance guarantee for the number of exemplars chosen by Algorithm 1. Here, we explain how such an improvement can be obtained.

Since Step 3 in Algorithm 1 uses an approximation algorithm for MSC, the performance guarantee with respect to the number of exemplars is $O(\log n)$, where $n = |X|$. Theoretically, one can get a better approximation by transforming the Exemplar Selection steps (i.e., Steps 2 and 3 of the algorithm) into that of finding a near-optimal dominating set for unit disk graphs in an Euclidean space whose dimension ℓ is the same as that of the points in X. This is done by placing an ℓ-dimensional ball of radius $\epsilon/2$ at each instance in X. The corresponding unit disk graph has a node for each instance in X and there is an edge between two nodes if the corresponding balls intersect or touch. It can be verified that any dominating set for this graph provides the necessary set of exemplars. An approximation scheme which provides a performance guarantee of $(1 + \delta)$ for any

fixed $\delta > 0$ is known for the minimum dominating set problem for such graphs [10]. Thus, one can obtain a performance guarantee of $(1 + \delta)$ for any fixed $\delta > 0$ with respect to the number of exemplars. However, this approximation scheme is impractical even for data sets of moderate size since its running time has the factor $O(n^{(1/\delta)^2})$. (Thus, even when $\delta = 0.5$, the running time has the factor $O(n^4)$.) For this reason, we decided to use the MSC-based approximation algorithm in our experiments.

11 Statement and Proof of Theorem 4.2

Statement of Theorem 4.2: The solution produced by Algorithm 2 satisfies the following properties: (i) The diameter of each cluster is at most $2(D^* + \epsilon)$, where D^* is the optimal diameter for a k-clustering of X and ϵ is the exemplar distance. (ii) The sets of exemplars for the k clusters are pairwise disjoint. (iii) The total number of instances with exemplars is at least $(1 - 1/\mathrm{e})Q^*$, where e is the base of the natural logarithm and Q^* is the maximum number of instances in X that can have exemplars under the condition that the total number of exemplars is at most β.

Proof: The proofs of Parts (i) and (ii) of this the theorem are identical to the ones given in the proof of Theorem 4.1. Part (iii) follows from the result from [12] that the greedy approximation algorithm for BMC covers at least $(1 - 1/\mathrm{e})Q^*$ elements, where Q^* is the maximum number of elements that can be covered using at most β sets. ■