Q-Means: A new Heuristic for Bi-Clustering

Yu WEI
Advanced Optimization lab
McMaster University
Supervised by Dr. Peng, J.
Joint work with Guoxuan Ma
Outline

- Introduction
- Motivation
- Reformulation and Q-Means Heuristic
- Generalized Q-Means
- Computational Results
- Conclusion and Future work
- References
Outline

- Introduction
- Motivation
- Reformulation and Q-Means Heuristic
- Generalized Q-Means
- Computational Results
- Conclusion and Future work
- Reference
Introduction

- Clustering analysis (unsupervised learning) in data mining:
  - Partitioning a given set of entities into several groups based on some similarity measurement

- Application:
  - marketing analysis
  - bioinformatics
  - pattern recognition etc.
Introduction

- Most of the clustering algorithms can be grouped into one of the following categories [1]:
  - Partitioning methods
  - Hierarchical methods
  - Density-based methods

- K-Means is the most popular clustering algorithm so far, and many partitioning algorithms are heuristics based on K-Means and its variants.
Clustering Analysis – Optimization Perspective

- Based on Minimum Squared Sum Criterion (MSSC).
  \[ \min J = \sum_{k=1}^{K} \sum_{i \in C_k} (s_i - c_k)^2 \]

- The objective function is nonconvex and nonsmooth [2]
  \[ \min_{c_1, \ldots, c_K} \sum_{i=1}^{n} \min\{ \|s_i - c_1\|^2, \ldots, \|s_i - c_K\|^2 \}. \]
K-Means Algorithm

- Initialization (choose initial cluster centers)
- Loop until termination condition is met:
  - 1. For each entity in data set, assign that entity to a cluster such that the distance from this entity to the center of that cluster is minimized.
  - 2. For each cluster, recalculate the means of the cluster based on the entities that belong to that cluster.
- end loop;
K-Means Algorithm - Illustration
K-Means Algorithm - Weaknesses

- Sensitive to initial choice of cluster center.

- Tend to converge to a local optimum rather than global optimum. How to get a ‘deep’ local optimum point with K-Means is still a very active research area.

- Can not handle constrained clustering problems
  - Size ratio Constraint.
  - Relational Constraints.
  - Optimization Constrains.
Outline

- Introduction
- Motivation
- Reformulation and Q-Means Heuristic
- Generalized Q-Means
- Computational Results
- Conclusion and Future work
- References
Motivation for Q-Means Algorithm

- Consider a bi-clustering problem.

\[
\min_{c_1, c_2} \sum_{i=1}^{n} \min\{\|s_i - c_1\|^2, \|s_i - c_2\|^2\}.
\]

- We have

\[
\exists t \in (0, 1), \quad s.t. \quad (1-t)c_1 + tc_2 = c.
\]

\[
c_2 = c_1 + Q(c - c_1). \quad Q = \frac{1}{t}
\]
Motivation for Q-Means Algorithm

- If Q is fixed, then the distance from the two cluster centers to the geometric center of the whole data set is fixed. This means that if the data set follows certain distribution, then the sizes of two clusters are also under control.

- This allows us to deal with the constrained clustering problem.

- In marketing analysis, sometimes the customers may want the size of clusters to be equal.
Outline

- Introduction
- Motivation
- Reformulation and Q-Means Heuristic
- Generalized Q-Means Heuristic
- Computational Results
- Conclusion and Future work
- References
Reformulation to QP

- Use the relation in bi-clustering, the original MSSC problem can be reformulated to a QP:

\[
\min_{c_1} \sum_{i=1}^{n} (\|s_i - c_1\|^2 + \min\{0, \phi_i(c_1)\}).
\]

- Define an active index set \( I = i : \phi_i(c_1) < 0 \),

\[
\min_{c_1} \sum_{i=1}^{n} \|s_i - c_1\|^2 + \sum_{i \in I} \phi_i(c_1).
\]
Q-Means Heuristic

Algorithm 2.1 Q-Means

Step 1. Initialization. Choose a point in dataset randomly as one of the cluster centers, say $C_1$, set $i = 1$.

Step 2. Iteration.

Step 2.1 Set the active set according to $C_i$, i.e. for every point $s_i$, check whether $\phi_i(c_i) \leq 0$.

Step 2.2. Solve the quadratic problem (11) to minimize the objective function according to current active index set, the solution is $C_{i+1}$. Goto Step 2 if $C_i \neq C_{i-1}$

Step 3. Rounding. Move some points from one cluster to the other, if not decrease, with least increase in obj function, such that the size ratio constrain is satisfied.
Q-Means Heuristic - Problems

- Still sensitive to initial point as K-Means:
Q-Means Heuristic - Problems
Q-Means Heuristic - Problems

Q: how could we make it less sensitive to initial starting points?

A: we can try for every entity in data set, the complexity is $O(n)$, if we do this with K-Means, it would be $O(n^2)$

However, what benefit could we get for enumerating all entities uas starting points in data set? Seems it guaranteed a ‘deep’ local optimum, but how ‘deep’? (Theoretical topics we working on)
Q-Means Heuristic - Revised

Algorithm 2.2 Q-Means

Step 1. Initialization. Choose every point in dataset as one of the cluster centers, and do step 2 once, keep the point that has minimal objective function, make it as $C_1$, set $i = 1$.

Step 2. Iteration.

Step 2.1. Set the active set according to $C_i$, i.e. for every point $s_i$, check whether $\phi_i(c_i) \leq 0$.

Step 2.2. Solve the quadratic problem (11) to minimize the objective function according to current active index set, the solution is $C_{i+1}$. Goto Step 2 if $C_i \neq C_{i-1}$.

Step 3. Rounding. Move some points from one cluster to the other, if not decrease, with least increase in obj function, such that the size ratio constrain is satisfied.
Outline

- Introduction
- Motivation
- Reformulation and Q-Means Heuristic
- Generalized Q-Means
- Computational Results
- Conclusion and Future work
- References
Generalized Q-Means

- Can we use Q-Means’ idea in unconstrained cluster problem?

- Q-Means provided a new approach to reduce the objective function value of MSSC problem, how could we reduce the objective function value more on the basis of the result of Q-Means?

- Relax Q after Q-Means stops.
Algorithm 2.4 Generalized Q-Means

*Step 1.* Choose a fix $Q$, do the 2-clustering using Q-Means

*Step 2.* Refine the result of Q-Means by conducting local heuristics, in particular, we use HK-Means (two phases) as our local heuristic, J-Means is a good choice as well. [5]
Outline

- Introduction
- Motivation
- Reformulation and Q-Means Heuristic
- Generalized Q-Means
- Computational Results
- Conclusion and Future work
- References
Synthetic Data Sets – Revised Q-Means

- Use the same data set that we illustrated sensitivity of Q-Means.
Synthetic Data Sets – Revised Q-Means

- One of the results from K-Means.
Synthetic Data Sets – Revised Q-Means

- Compare of objective function value.

<table>
<thead>
<tr>
<th>Method</th>
<th>Value</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Means (Best)</td>
<td>1973.778934</td>
<td>2729.907621</td>
</tr>
<tr>
<td>K-Means (Worst)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Synthetic Data Sets – Generalized Q-Means

- We used a random generator to produce various two-dimensional synthetic data sets approximately in the mixture Gaussian distribution.

- S01 $v = 0.05$, $r = 0.4$, S02 $v = 0.10$, $r = 0.4$. S03 $v = 0.15$, $r = 0.4$
Synthetic Data Sets - Results

<table>
<thead>
<tr>
<th>data set</th>
<th>average obj value (K-Means)</th>
<th>obj value (Q-Means)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01</td>
<td>273.087161</td>
<td>273.086971</td>
</tr>
<tr>
<td>S02</td>
<td>269.384920</td>
<td>269.384890</td>
</tr>
<tr>
<td>S03</td>
<td>275.514638</td>
<td>275.514575</td>
</tr>
</tbody>
</table>

*Note:* The best obj value is obtained from K-Means is the same with Q-Means. But the difference here is we just run Q-Means once.
Real-life Data Sets

- We test Q-Means on some standard tests in clustering literature as well. For Soybean data (small) and Soybean data (large) from the UCI Machine Learning Repository, the result is:

<table>
<thead>
<tr>
<th>data set</th>
<th>average obj value (K-Means)</th>
<th>obj value (Q-Means)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soybean(S)</td>
<td>527.791892</td>
<td>404.459259</td>
</tr>
<tr>
<td>Soybean(L)</td>
<td>4489.003533</td>
<td>4466.866548</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- Motivation
- Reformulation and Q-Means Heuristic
- Generalized Q-Means
- Computational Results
- Conclusion and Future work
- References
Conclusion

(1) For fixed size ratio problem, Q-Means performs fairly good.

(2) Now, Q-Means can only deal with bi-clustering, For $K > 2$, one can do a hierarchical divisive clustering where each step using the bi-clustering.
Future Work

- Prove the approximate rate (how ‘deep’) we got from enumerating all entities in data set. This is one of the important differences from K-Means.

- Avoid local optimum in Q-Means. If we do not enumerate all entities, we can use some techniques as used in K-Means to avoid local optimum.
References


Questions?
Thanks! 😊