Retrieval of Optimal Subspace Clusters Set for an Effective Similarity Search in a High Dimensional Spaces

Ivan Sudos
SPBU
Retrieval problem

Time series

Gene expressions

Media collections

Vector space; d >= 10
Retrieval problem

(0,4,10,15,0,-1,1,3,16,10,1,-2)
(0,2,-1,-5,0,-1,2,6,11,-1,4,2)
(0,4,-5,-3,0,-2,1,1,2,7,19,21)
(0,1,6,4,9,-4,-2,-5,-3,7,15,5)
....
(0,2,-1,-5,1,1,2,6,10,-1,4,3)
....
(1,1,-3,-5,1,1,2,7,10,-1,4,3)
....
(2,0,-1,-4,3,6,2,7,10,-1,4,3)

q = (0,2,-1,-5,0,-1,3,7,12,0,2,2)
Curse of dimensionality

- Near equidistant in terms of Euclid distance
- Space volume grows exponentially
- Number of attributes under analysis is high
Inapplicability of low-dimensional indexing techniques

• Space partitioning index structures (like R-trees) lead to exponentially explosion of index size.

• Random projection leads to inability of relevant nearest neighbour detection
Clustering approach to indexing

- Clustering is a convenient form of unsupervised learning;
- Cluster has a role of vector-approximation file;
- Retrieval task is reduced to classification.
Subspace and projection clustering

- An approach to fight curse of dimensionality for clustering
- Particular data points form clusters in a particular subspace

Subspace Cluster = \{S,C\};
where S — subset of dimensions;
C- subset of data vectors
Retrieval approach with subspace clusters

- Calculate an approximate relevance of the cluster with respect to a given query
- Select the set of the most relevant clusters
- Continue refinements
Problem: curse of dimensionality again

• The number of all clusters in all possible subspaces can be significantly high

• Cluster in subspace of dimensionality > 5 is still complex to analyse; Q becomes complex;
Problem: what we can do

Consider only the best (optimal) subset of clusters as an index set.
Optimization problem

\[
E(R(c^*, q)) \rightarrow \text{max}, \quad \text{where} \quad c^* = \arg \max_{c \in C} R_{\text{approx}}(c, q)
\]

\[
\begin{align*}
N \leq N_{\text{max}} \\
E(g_q(|V|, |s|)) &\leq \gamma \forall q
\end{align*}
\]
Solution

• Hard to solve in a general way; Say, q distributed uniformly, then we need to calculate complex multiple integral sums;

• We will attempt to reduce the problem to a knapsack problem

• We need weights and values;
Solution: Knapsack problem reduction

• Weight of the cluster can be assigned to 1
  \[ w = 1 \]

• To introduce value function an estimation algorithm was invented;

\[
(u(x), w(x)) := (R(x, c_i); \ R_{approx}(x, c_i))
\]

\[
v_i = corr(u, w)
\]
Solution: Results

- Weather observation data
- ~17 millions of vectors
- Dimensionality 200
- 1630 clusters detected by MAFIA
- Average dimensionality 9.5
Solution: Results

- **Relevance**
  - Graph showing relevance decreasing as a function of some parameter.

- **Latency**
  - Graph showing latency decreasing as a function of some parameter.
Q & A