Concentration Effect: A Motivation of High-Dimensional Outlier Detection

Presenter: Jaehyeong Ahn

Konkuk University

jayahn0104@gmail.com

February 19, 2021

Table of Contents

Introduction

Introduction

- (Unsupervised) Outlier Detection techniques generally measure the degree of outlierness by computing the distances between data points in (full dimensional) feature space
- High-dimensional data causes a “concentration effect” which makes the distances between all data points become similar (beyer et. al. 1999). This phenomenon becomes the motivational problem of outlier detection in high-dimensional data
- In this presentation, we investigate “concentration effect” and some cases that can prevent this phenomenon

Concentration Effect

Concentration Effect

Definition 1

- m: the number of dimensions
- F_{data_m}: an infinite sequence of data distributions, $m = 1, 2, \cdots$
  - $\boldsymbol{X}^{(m)} \sim F_{data_m}$: an arbitrary random vector distributed as F_{data_m}
  - $\boldsymbol{x}_1^{(m)}, \cdots, \boldsymbol{x}_n^{(m)} \sim F_{data_m}$: n independent data points per m
- F_{query_m}: an infinite sequence of query distributions, $m = 1, 2, \cdots$
  - $\boldsymbol{Q}^{(m)} \sim F_{query_m}$: a query point chosen independently from $\boldsymbol{x}_i^{(m)}, \forall_i$
- $d_{m,p}(\boldsymbol{X}^{(m)}, \boldsymbol{Q}^{(m)})$: the function gives the L_p distance between $\boldsymbol{X}^{(m)}$ and $\boldsymbol{Q}^{(m)}$, $\forall p > 0$
- $\mathrm{DMAX}^{(m)} = \max\{\mathrm{d}_{m,p}(\boldsymbol{x}_i^{(m)}, \boldsymbol{Q}^{(m)}) | 1 \leq i \leq n\}$
- $\mathrm{DMIN}^{(m)} = \min\{\mathrm{d}_{m,p}(\boldsymbol{x}_i^{(m)}, \boldsymbol{Q}^{(m)}) | 1 \leq i \leq n\}$

Concentration Effect

- Below theorem states that assuming the distance distribution behaves a certain way as m increases, *the difference in distance between the query point and all data points becomes "negligible"*

Theorem 1 (*ConcentrationEffect*)

Under the certain conditions in Definition 1,

If $\lim_{m\to\infty} \mathrm{Var}\left(\frac{\mathrm{d_{m,p}}(\boldsymbol{X}^{(\mathrm{m})},\boldsymbol{Q}^{(\mathrm{m})})}{\mathrm{E}[\mathrm{d_{m,p}}(\boldsymbol{X}^{(\mathrm{m})},\boldsymbol{Q}^{(\mathrm{m})})]}\right) = 0$, Then $\frac{\mathrm{DMAX}^{(\mathrm{m})}}{\mathrm{DMIN}^{(\mathrm{m})}} \to_p 1$

  - Where the operators $\mathrm{E}[\cdot]$ and $\mathrm{Var}[\cdot]$ refer to the theoretical expectation and variance of the distribution of the random variable $d_{m,p}(\boldsymbol{X}^{(m)}, \boldsymbol{Q}^{(m)})$
  - It is assumed that $\mathrm{E}[\mathrm{d_{m,p}}(\boldsymbol{X}^{(\mathrm{m})}, \boldsymbol{Q}^{(\mathrm{m})})]$ and $\mathrm{Var}(\mathrm{d_{m,p}}(\boldsymbol{X}^{(\mathrm{m})}, \boldsymbol{Q}^{(\mathrm{m})}))$ are finite and $\mathrm{E}[\mathrm{d_{m,p}}(\boldsymbol{X}^{(\mathrm{m})}, \boldsymbol{Q}^{(\mathrm{m})})] \neq 0$

Proof of Theorem 1

- Some Results from Probability Theory

Lemma 1

If $B_1, B_2, \cdots$ is a sequence of random variables with finite variance and $\lim_{m\to\infty} \mathrm{E[B_m] = b}$ and $\lim_{m\to\infty} \mathrm{Var(B_m)} = 0$ then $B_m \to_p b$

A version of Slutsky's theorem

Let $\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots$ be random variables (or vectors) and g be a continuous function. If $\boldsymbol{A}_m \to_p \boldsymbol{c}$ and $g(\boldsymbol{c})$ is finite then $g(\boldsymbol{A}_m) \to_p g(\boldsymbol{c})$

Corollary 1

If $X_1, X_2, \cdots$ and $Y_1, Y_2, \cdots$ are sequences or random variables s.t. $X_m \to_p a$ and $Y_m \to_p b \neq 0$ then $X_m / Y_m \to_p a/b$.

Proof of Theorem 1

- Let
  - $\mu^{(m)} = \mathrm{E}[\mathrm{d}_{\mathrm{m,p}}(\boldsymbol{X}^{(\mathrm{m})}, \boldsymbol{Q}^{(\mathrm{m})})]$
  - $V^{(m)} = d_{m,p}(\boldsymbol{X}^{(m)}, \boldsymbol{Q}^{(m)})/\mu^{(m)}$

- **Part 1**: We will show that $V^{(m)} \to_p 1$
  - $\lim_{m\to\infty} \mathrm{E}[\mathrm{V}^{(\mathrm{m})}] = \lim_{\mathrm{m}\to\infty} \mathrm{E}[\mathrm{d}_{\mathrm{m,p}}(\boldsymbol{X}^{(\mathrm{m})}, \boldsymbol{Q}^{(\mathrm{m})})]/\mu^{(\mathrm{m})} = 1$
  - $\lim_{m\to\infty} \mathrm{Var}(\mathrm{V}^{(\mathrm{m})}) =0$ (By the condition of the theorem)

  $\Rightarrow$ Based on Lemma 1, we can conclude that $V^{(m)} \to_p 1$

Proof of Theorem 1

- **Part 2**: We'll show that if $V^{(m)} \rightarrow_p 1$ then $\frac{\mathrm{DMAX^{(m)}}}{\mathrm{DMIN^{(m)}}} \rightarrow_p 1$
  - Let $\boldsymbol{W}^{(m)} = \left(d_{m,p}(\boldsymbol{x}_1^{(m)}, \boldsymbol{Q}^{(m)})/\mu^{(m)}, \cdots, d_{m,p}(\boldsymbol{x}_n^{(m)}, \boldsymbol{Q}^{(m)})/\mu^{(m)} \right)$
  - Since each element of the vector $\boldsymbol{W}^{(m)}$ has the same distribution as $V^{(m)}$, it follows that $\boldsymbol{W}^{(m)} \rightarrow_p (1, \cdots, 1)$
  - By using the Slutsky's theorem, we can conclude that $\min(\boldsymbol{W}^{(m)}) \rightarrow_p \min(1, \cdots, 1) = 1$ and $\max(\boldsymbol{W}^{(m)}) \rightarrow_p \max(1, \cdots, 1) = 1$
  - Using Corollary 1 on $\max(\boldsymbol{W}^{(m)})$ and $\min(\boldsymbol{W}^{(m)})$ we get

$$\frac{\max(\boldsymbol{W}^{(m)})}{\min(\boldsymbol{W}^{(m)})} \rightarrow_p \frac{1}{1} = 1$$

  - Note that $\mathrm{DMAX^{(m)}} = \mu^{(m)} \max(\boldsymbol{W}^{(m)})$ and $\mathrm{DMIN^{(m)}} = \mu^{(m)} \min(\boldsymbol{W}^{(m)})$

$$\frac{\mathrm{DMAX^{(m)}}}{\mathrm{DMIN^{(m)}}} = \frac{\mu^{(m)} \max(\boldsymbol{W}^{(m)})}{\mu^{(m)} \min(\boldsymbol{W}^{(m)})} = \frac{\max(\boldsymbol{W}^{(m)})}{\min(\boldsymbol{W}^{(m)})} \rightarrow_p 1$$

Concentration Effect

Immediate Questions

- When does the condition of Theorem 1 hold?
  (i.e. $\lim_{m\to\infty} \mathrm{Var}\left(\frac{\mathrm{d_{m,p}}(\boldsymbol{X}^{(\mathrm{m})},\boldsymbol{Q}^{(\mathrm{m})})}{\mathrm{E}[\mathrm{d_{m,p}}(\boldsymbol{X}^{(\mathrm{m})},\boldsymbol{Q}^{(\mathrm{m})})]}\right) = \lim_{\mathrm{m}\to\infty} \frac{\mathrm{Var}(\mathrm{d_{m,p}}(\boldsymbol{X}^{(\mathrm{m})},\boldsymbol{Q}^{(\mathrm{m})}))}{(\mathrm{E}[\mathrm{d_{m,p}}(\boldsymbol{X}^{(\mathrm{m})},\boldsymbol{Q}^{(\mathrm{m})})])^2} = 0$)
  - We will provide some scenarios that do and do not satisfy the condition
- For situations in which the condition is satisfied, "at what rate" do distances between points become indistinct as dimensionality increases?
  - This issue is more difficult to tackle analytically. Therefore a set of simulations will be provided

Applicability of Concentration Effect

Example 1. *IID Dimensions with Query and Data Independence*

- Assumptions
  - The data distribution and query distribution are IID in all dimensions
  - The query point is chosen independently of the data points
- Proof

$$\lim_{m\to\infty} \frac{\mathrm{Var}(\sum_{j=1}^{m} |X_j - Q_j|^p)}{(E[(\sum_{j=1}^{m} |X_j - Q_j|^p)])^2} = \lim_{m\to\infty} \frac{m\sigma^2}{m^2\mu^2} = 0$$

  - Recall that L_p distance function $d_{m,p}(\boldsymbol{X}, \boldsymbol{Q}) = \sum_{j=1}^{m} |X_j - Q_j|^p$
  - Note that identical per dimension characteristics in our assumptions allow us to say that for j, $|X_j - Q_j|^p$ is some random variable U_j, and all U_j's are IID with mean μ and σ^2
  - Thus $E[\sum_{j=1}^{m} U_j] = m\mu$ and $\mathrm{Var}(\sum_{j=1}^{m} U_j) = \sum_{j=1}^{m} \mathrm{Var}(U_j) = m\sigma^2$

Applicability of Concentration Effect

Example 2. *Identical Dimensions with no Independence*

- Assumption
  - All dimensions of both the query point and the data points follow identical distributions, but are completely dependent (i.e., value for dimension 1 = value for dimension 2 = $\cdots$)
- Proof

$$\lim_{m\to\infty} \frac{\mathrm{Var}(\sum_{j=1}^{m} |X_j - Q_j|^p)}{(E[(\sum_{j=1}^{m} |X_j - Q_j|^p)])^2} = \lim_{m\to\infty} \frac{m^2\sigma^2}{m^2\mu^2} = \frac{\sigma^2}{\mu^2} \neq 0$$

  - $E[|X_j - Q_j|^p] = m\mu$
    - Since Expected values are not affected by dependence
  - $\mathrm{Var}(\sum_{j=1}^{m} |X_j - Q_j|^p) = \mathrm{Var}(m|X_j - Q_j|^p) = m^2\mathrm{Var}(|X_j - Q_j|^p) = m^2\sigma^2$
    - Since all dimensions are correlated (dependent) the sum is performed inside the variance

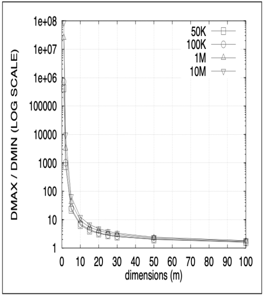
Applicability of Concentration Effect

Example 3. *Marginal Data and Query Distributions Change with Dimensionality*

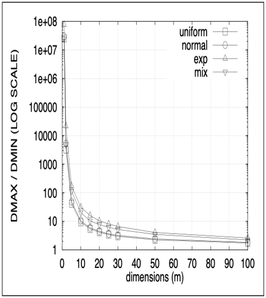
- Assumptions
  - The marginal distributions of data and queries change with dimensionality (not identical)
- Proof
  - Even in this case, the condition of Theorem 1 is satisfied
  - We will show it by some empirical experiments

Applicability of Concentration Effect

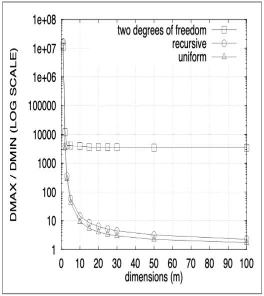
Empirical Results



(a) Size varies - Uniform distribution



(b) Distribution varies - 1M samples



(c) Correlated dimensions - 1M samples

Figure: These figures show the relationship between dimensionality and $\mathrm{DMAX_m}/\mathrm{DMIN_m}$ with various sample size and distributions

Empirical Results

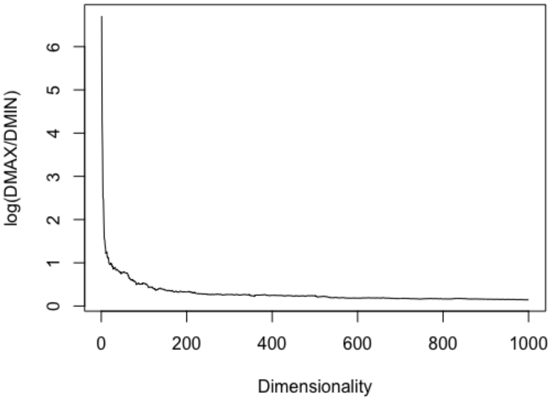


Figure: Standard normal distribution - 1000 samples, 1000 dimensions

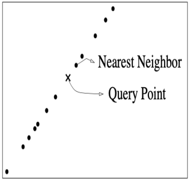
Applicability of Concentration Effect

- These results state that under broad conditions (broader than the IID dimensions assumption which other work assumes) concentration phenomenon shows up
- And the distinction between nearest and farthest neighbors may blur with as few as 15 dimensions
- Hence, under certain broad conditions, "Nearest Neighbor" in high-dimensional data becomes meaningless which means contrast in distances to different data points becomes non-existent
- However it does not mean that high-dimensional Nearest Neighbor is never meaningful. We will provide some typical cases which make "Nearest Neighbor" still meaningful

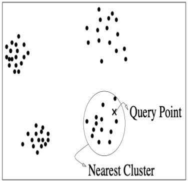
Generalized Cases that Prevent Concentration Effect

Generalized Cases that Prevent Concentration Effect

- **Low Intrinsic Structure**
  - Durrant and Kaban (2009) generalize the **Example 2** and show that when there exist richness of correlations between the variables the concentration phenomenon would not appear by using the latent variable model
- **Separable Cluster Structure**
  - Bennett et. al. (1999) show that when there exist separable cluster structure, the distance between two data points in different clusters (between cluster distance) dominates the distance between two points in the same cluster (within cluster distance). Thus concentration effect does not show up for between cluster distance



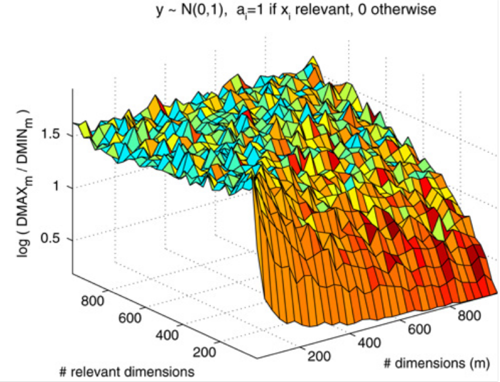
(a) Low Intrinsic Structure



(b) Separable Cluster Structure

Generalized Cases that Prevent Concentration Effect

- Durrant and Kaban (2004) defines the "relevant dimension" as the variable which has correlation with other variables in the perspective of the latent variable model
- Below Results show that the proportion of relevant dimensions among all dimensions plays a key role for preventing the concentration effect



Conclusion

Conclusion

- We illustrated the phenomenon called "concentration effect" which makes the distances of all data points become similar
- We showed that this phenomenon shows up in broad certain conditions (broader that IID dimensions) and the distinction between nearest and farthest neighbors may blur with as few as 15 dimensions
- There were two generalized cases that can prevent the phenomenon which have some (continuous/discrete) latent structure in the data
- These results lead to a statement that Outlier detection in high-dimensional data should carefully consider this phenomenon
- Subspace outlier detection is a representative technique for high-dimensional outlier detection. This method selects some subspaces (subset of variables) and then measure the outlierness on the specific subspace

References

- Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U. (1999, January). When is "nearest neighbor" meaningful?. In International conference on database theory (pp. 217-235). Springer, Berlin, Heidelberg.
- Durrant, R. J., Kabán, A. (2009). When is 'nearest neighbour'meaningful: A converse theorem and implications. Journal of Complexity, 25(4), 385-397.
- Bennett, K. P., Fayyad, U., Geiger, D. (1999, August). Density-based indexing for approximate nearest-neighbor queries. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 233-243).

Q&A