Density-Based Clustering Validation: Supplementary Material

Davoud Moulavi* Pablo A. Jaskowiak*† Ricardo J. G. B. Campello† Arthur Zimek‡
Jörg Sander*

Supplementing the article “Density-Based Clustering Validation” by the authors, published at 2014 SIAM International Conference on Data Mining (SDM), 2014, Philadelphia, PA, USA

1 Proofs

Proposition 3.1. *The all points core distance of each object* $\mathbf{o}$ $a_{pts}coredist(\mathbf{o})$ *is between the second and last nearest neighbor distance of that object, i.e.,*

$$KNN(\mathbf{o},2) \le a_{pts}coredist(\mathbf{o}) \le KNN(\mathbf{o},n)$$

Proof. We have $\forall 2 \le i \le n$,

$$\begin{aligned} KNN(\mathbf{o},2) &\le KNN(\mathbf{o},i) \Rightarrow \\ \left(\frac{1}{KNN(\mathbf{o},i)}\right)^d &\le \left(\frac{1}{KNN(\mathbf{o},2)}\right)^d \end{aligned}$$

Therefore,

$$\begin{aligned} \sum_{i=2}^{n}\left(\frac{1}{KNN\left(\mathbf{o},i\right)}\right)^d &\le (n-1)\left(\frac{1}{KNN\left(\mathbf{o},2\right)}\right)^d \\ \left((n-1)\frac{\frac{1}{KNN(\mathbf{o},2)}^d}{n-1}\right)^{-\frac{1}{d}} &\le \left(\frac{\sum_{i=2}^{n}\left(\frac{1}{KNN(\mathbf{o},i)}\right)^d}{n-1}\right)^{-\frac{1}{d}} \end{aligned}$$

leading to Equation 1.

$$(1) \qquad KNN(\mathbf{o},2) \le a_{pts}coredist(\mathbf{o}).$$

The upper bound inequality can be similarly proved. Note that the above proof is valid for both Euclidien and Squared Euclidean distance dissimilarity measures.

Proposition 3.2. *Let* n *objects be uniformly distributed random variables in a* d*-dimensional unit hypersphere and* $\mathbf{o}$ *be an object in the center of this hypersphere. With Euclidean distance as dissimilarity measure the core distance of* $\mathbf{o}$ *is:*

$$(2) \quad a_{pts}coredist(\mathbf{o}) \approx (\ln(n-1)+\gamma+\epsilon)^{-\frac{1}{d}} \approx \ln(n)^{-\frac{1}{d}}$$

where $\gamma \approx 0.5772$ *is the so-called Euler-Mascheroni constant and* $\epsilon = \frac{1}{2n} - \sum_1^\infty \frac{B_{2k}}{2kn^{2k}} \approx \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4}$ *where* B_k *is the Bernoulli number.*

Proof. First we prove the following:

$$(3) \qquad \forall 1 \le i \le j \le n, \frac{KNN(\mathbf{o},i)}{KNN(\mathbf{o},j)} \approx \left(\frac{i-1}{j-1}\right)^{\frac{1}{d}}$$

$KNN(\mathbf{o},i)$ being the expected value of the i^{th} nearest neighbor of the center object $\mathbf{o}$ in the hypersphere.

The complete proof needs lengthy detailed discussion. We provide a sketch of the proof. The average minimum volume of the hypersphere that contains the i^{th} nearest neighbor of object $\mathbf{o}$ (center of the hypersphere) is given by the following:

$$V_i \approx \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}r_i^d$$

where V_i is the minimum volume hypersphere with radius r_i that contains $KNN(\mathbf{o},i)$. Therefore we have $KNN(\mathbf{o},i) = r_i$. Also in uniform distribution the ratio of the expected volumes of the hyperspheres that contain the i^{th} and j^{th} nearest neighbors to object $\mathbf{o}$ is $\frac{V_i}{V_j} \approx \frac{i-1}{j-1}$. Therefore,

$$(4) \qquad \frac{V_i}{V_j} \approx \left(\frac{r_i}{r_j}\right)^d$$

$$(5) \qquad \frac{r_i}{r_j} = \frac{KNN(\mathbf{o},i)}{KNN(\mathbf{o},j)} \approx \left(\frac{i-1}{j-1}\right)^{\frac{1}{d}}$$

From Equation (3) we have:

$$\frac{1}{KNN(\mathbf{o},i)} = \left(\frac{n-1}{i-1}\right)^{\frac{1}{d}}\frac{1}{KNN(\mathbf{o},n)}$$

*Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada, {moulavi, jaskowia, jsander}@ualberta.ca

†Department of Computer Science, University of São Paulo, São Carlos, Brazil, {pablo, campello}icmc.usp.br

‡Ludwig-Maximilians-Universität München, Munich, Germany, zimek@dbs.ifi.lmu.de

Therefore,

$$a_{pts}\,\mathit{coredist}(\mathbf{o}) \quad = \left(\frac{\sum_{i=2}^{n} \left(\frac{n-1}{i-1}^{\frac{1}{d}} \frac{1}{KNN(\mathbf{o},n)} \right)^{d}}{n-1} \right)^{-\frac{1}{d}}$$

we have $KNN(\mathbf{o}, n) \approx 1$ therefore:

$$\begin{aligned}
a_{pts}\,\mathit{coredist}(\mathbf{o}) \quad &= \left(\frac{\sum_{i=2}^{n} \left(\frac{n-1}{i-1}^{\frac{1}{d}} \right)^{d}}{n-1} \right)^{-\frac{1}{d}} \\
&= \left(\frac{\sum_{i=2}^{n} \left(\frac{n-1}{i-1} \right)}{n-1} \right)^{-\frac{1}{d}} \\
&= \left((n-1) \frac{\sum_{i=2}^{n} \left(\frac{1}{i-1} \right)}{n-1} \right)^{-\frac{1}{d}} \\
&= \left(\sum_{i=2}^{n} \left(\frac{1}{i-1} \right) \right)^{-\frac{1}{d}} \\
&= (\ln(n-1) + \gamma + \epsilon)^{-\frac{1}{d}} \approx \ln(n)^{-\frac{1}{d}}
\end{aligned}$$

$\sum_{i=1}^{n} \left(\frac{1}{i} \right)$ is called the harmonic series.

Proposition 3.3. *For $a_{pts}\,\mathit{coredist}(\mathbf{o})$ (Proposition 3.2), we have:*

$$(6) \qquad a_{pts}\,\mathit{coredist}(\mathbf{o}) \approx \ln(n)^{-\frac{1}{d}} \approx KNN(\mathbf{o}, j),$$

with j being the closest natural number to $\frac{n}{\ln(n)}$.

Proof. We know that $KNN(\mathbf{o}, n) \approx 1$, therefore, considering Equation (3), we have:

$$\begin{aligned}
\frac{KNN\left(\mathbf{o}, \frac{n}{\ln(n)}\right)}{KNN(\mathbf{o}, n)} \quad &\approx \quad \left(\frac{\frac{n}{\ln(n)} - 1}{n-1} \right)^{\frac{1}{d}} \\
&\approx \quad \left(\frac{\frac{n}{\ln(n)}}{n} \right)^{\frac{1}{d}} \\
&\approx \quad \ln(n)^{-\frac{1}{d}} \\
a_{pts}\,\mathit{coredist}(\mathbf{o}) \quad &\approx \quad KNN(\mathbf{o}, j)
\end{aligned}$$

This proposition shows that the core distance of the object $\mathbf{o}$ is approximately equal to the same nearest neighbor distance of $\mathbf{o}$ independent of the dimensionality of the data space.

Proposition 3.4. *If the dissimilarity measure in Proposition 3.2 is squared Euclidean distance the core distance of $\mathbf{o}$ is:*

$$(7) \qquad a_{pts}\,\mathit{coredist}(\mathbf{o}) \approx (1.65 * n)^{-\frac{1}{d}}$$

Proof. If the dissimilarity measure is squared Euclidean distance, we have:

$$(8) \qquad \forall 1 \leq i \leq j \leq n, \frac{KNN(\mathbf{o}, i)}{KNN(\mathbf{o}, j)} \approx \left(\frac{i-1}{j-1} \right)^{\frac{2}{d}}$$

Therefore:

$$\frac{1}{KNN(\mathbf{o}, i)} = \left(\frac{n-1}{i-1} \right)^{\frac{2}{d}} \frac{1}{KNN(\mathbf{o}, n)}$$

Thus,

$$a_{pts}\,\mathit{coredist}(\mathbf{o}) \quad = \left(\frac{\sum_{i=2}^{n} \left(\frac{n-1}{i-1}^{\frac{2}{d}} \frac{1}{KNN(\mathbf{o},n)} \right)^{d}}{n-1} \right)^{-\frac{1}{d}}$$

we have $KNN(\mathbf{o}, n) \approx 1$ therefore:

$$\begin{aligned}
a_{pts}\,\mathit{coredist}(\mathbf{o}) \quad &= \left(\frac{\sum_{i=2}^{n} \left(\frac{n-1}{i-1}^{\frac{2}{d}} \right)^{d}}{n-1} \right)^{-\frac{1}{d}} \\
&= \left(\frac{\sum_{i=2}^{n} \left(\frac{n-1}{i-1} \right)^{2}}{n-1} \right)^{-\frac{1}{d}} \\
&= \left((n-1)^{2} \frac{\sum_{i=2}^{n} \left(\frac{1}{i-1} \right)^{2}}{n-1} \right)^{-\frac{1}{d}} \\
&= \left((n-1) \sum_{i=2}^{n} \left(\frac{1}{i-1} \right)^{2} \right)^{-\frac{1}{d}} \\
&\approx (1.65 * n)^{-\frac{1}{d}}
\end{aligned}$$

using well-known Basel problem it can be easily proved that:

$$1 \leq \sum_{i=1}^{n} \left(\frac{1}{i} \right)^{2} < 1.65$$

Proposition 3.5. *For $a_{pts}\,\mathit{coredist}(\mathbf{o})$ (Proposition 3.4), we have:*

$$(9) \qquad a_{pts}\,\mathit{coredist}(\mathbf{o}) \approx (1.65 * n)^{-\frac{1}{d}} \approx KNN(\mathbf{o}, j),$$

with j being the closest natural number to $\sqrt(n/1.65)$.

Proof. We know that $KNN(\mathbf{o}, n) \approx 1$, therefore, considering Equation (8), we have:

$$
\begin{aligned}
\frac{KNN\left(\mathbf{o}, \sqrt{\frac{n}{1.65}}\right)}{KNN(\mathbf{o}, n)} &\approx \left(\frac{\sqrt{\frac{n}{1.65}} - 1}{n-1}\right)^{\frac{2}{d}} \\
&\approx \left(\frac{\sqrt{\frac{n}{1.65}}}{n}\right)^{\frac{2}{d}} \\
&\approx (1.65 * n)^{-\frac{1}{d}} \\
a_{pts}\, coredist(\mathbf{o}) &\approx KNN(\mathbf{o}, j)
\end{aligned}
$$

Similar to Proposition 3.3 this proposition also shows that the core distance of the object **o** is approximately equal to the same nearest neighbor distance of **o** independent of the dimensionality of the data space. Comparing propositions 3.3 and 3.5 confirms that by applying squared Euclidean distance the effect of the first property becomes stronger and the core distance represents smaller neighborhood of the objects.

2 Comments on Noise

Here we show that the weighted averaging approach from DBCV, as shown in Equation 3.5, is exactly the same as penalizing the other relative validity measures based on the proportion of the noise objects in the dataset. Let $|N|$ be the cardinality of noise objects and $|O|$ be the cardinality of all objects in the dataset.

$$
\begin{aligned}
DBCV(C) &= \textstyle\sum_{i=1}^{i=l} \frac{|C_i|}{|O|} V_C(C_i) \\
&= \frac{|O|-|N|}{|O|} \textstyle\sum_{i=1}^{i=l} \frac{|C_i|}{|O|-|N|} V_C(C_i)
\end{aligned} \tag{10}
$$

Note that $\sum_{i=1}^{i=l} \frac{|C_i|}{|O|-|N|} V_C(C_i)$ is equal to removing the noise objects and calculating DBCV, whereas $\frac{|O|-|N|}{|O|}$ is penalizing the resulting value proportional to the number of noise objects that are left out from the partition.