PERSIST: A New Probabilistic Model For Data Denoising

Sergiu Scrob

Abstract

The paper proposes a more efficient solution for reducing the data noise level, using probability estimation for radius-based spatial inference and sampling techniques.

Keywords: data denoising, probabilistic model, radius-based spatial inference, spatial probability.

1 Introduction

The majority of clustering and classification models from scientific library like **Scikit-learn**, fails to satisfactorily denoise the data shown in Figure 1. The dataset consists of 50 subsets of 100k points positioned inside and outside the figure, forming two classes of points.

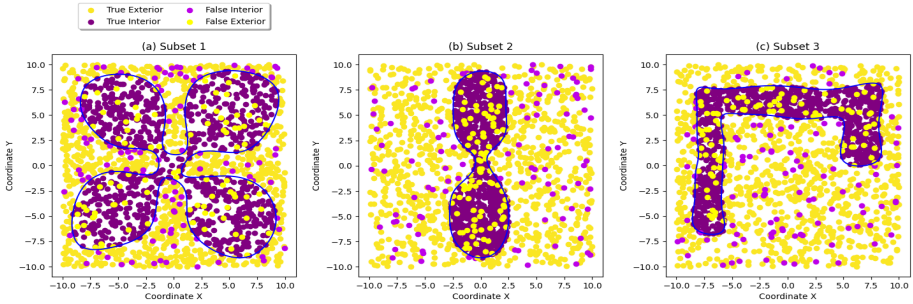


Figure 1. Examples of the first three subsets from the dataset

©2024 by Sergiu Scrob

2 Defining a new denoising model

The new probabilistic model named PERSIST: Probability Estimation for Radius-based Spatial Inference and Sampling Techniques, aims to reduce the data noise level by using the concepts of spatial and frequency probabilities, which represents the proximity and density of the points in the vicinity of the centroid.

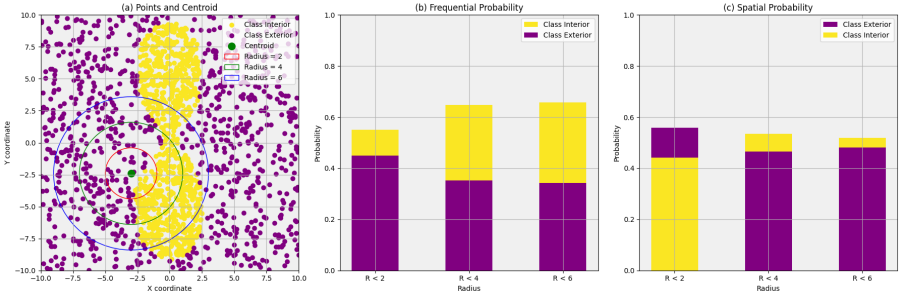


Figure 2. Local Probability Distributions by Centroid's Radius

2.1 The mathematical model

Let: $P = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}, n \in \mathbb{N}$; $\mathbf{P_I} = \{(x_i, y_i) \mid (x_i, y_i) \in P, label_i = 1\}$; $\mathbf{P_E} = \{(x_i, y_i) \mid (x_i, y_i) \in P, label_i = 2\}$; $\mathbf{r}$ is the radius from centroid; $c = \{c_x, c_y\}$ as centroid point, $p = \{p_x, p_y\}$ as any point, where $c, p \in P$. The *proximity*(c, P, r) as $\mathbb{R}^2 \times \mathcal{P}(\mathbb{R}^2) \times \mathbb{R} \to \mathcal{P}(\mathbb{R}^2) : \mathcal{N} \leftarrow \mathcal{N} \cup \mathcal{N}'$ where $\mathcal{P}(\mathbb{R}^2)$ is the power set of $\mathbb{R}^2$ and $\mathcal{N}'$ is the set of points unique by their position, defined as:

$$\begin{cases} p_i \mid d(\mathbf{c}, p_i) \leq r \wedge (d_i, \theta_i, \phi_i) \notin \mathcal{U}, \mathcal{U} \leftarrow \mathcal{U} \cup \{(d_i, \theta_i, \phi_i)\} \\ \emptyset, \text{if } \forall (d_j, \theta_j, \phi_j) \in \mathcal{U}, |d_i - d_j| \geq \alpha \wedge |\theta_i - \theta_j| \geq \beta \wedge |\phi_i - \phi_j| \geq \gamma \end{cases} \tag{1}$$

$i = (i_1, \ldots, i_n), j = (j_1, \ldots, j_k), i, j, n, k \in \mathbb{Z}^+ n \leq |P|, k \leq |\mathcal{U}|$, d_i is Euclidean distance, θ_i is azimuthal angle [**ZHANG20211333**], ϕ_i is polar angle.

$$d(\mathbf{p}, \mathbf{c}) = \sqrt{(p_x - c_x)^2 + (p_y - c_y)^2} \tag{2}$$

$$\theta(\mathbf{p}, \mathbf{c}) = \arccos\left(\frac{p_y - c_y + \epsilon}{d(\mathbf{p}, \mathbf{c}) + \epsilon}\right) \times \left(\frac{180}{\pi}\right), \text{ if } 0 < \epsilon < 0.001 \tag{3}$$

$$\phi(\mathbf{p}, \mathbf{c}) = \arctan 2\,(p_x - c_x, p_y - c_y) \times \left(\frac{180}{\pi}\right) \tag{4}$$

Let $P_{ri} = \textit{proximity}(\mathbf{c}, \mathbf{P_I}, r)$, $P_{re} = \textit{proximity}(\mathbf{c}, \mathbf{P_E}, r)$, $P_{re} \subset \mathbf{P_E}$, $P_{ri} \subset \mathbf{P_I}$

$$\text{Retrieve}(c, P_I, P_E, r) = \begin{cases} \text{Retrieve}(c, P_I, P_E, r + \Delta), \text{if } |P_{ri}| = |P_{re}| \\ (|P_{ri}|, |P_{re}|)\,, \text{otherwise} \end{cases} \tag{5}$$

$$\text{where } \Delta = \begin{cases} 0.369, & \text{if } |P_{ri}| + |P_{re}| > 3 \\ 0.123, & \text{otherwise} \end{cases}$$

Let $\mathbf{D}_i = \{d(\mathbf{c}, \mathbf{p}) \mid \mathbf{p} \in \mathbf{P}_{ri}\}$ and $\mathbf{D}_e = \{d(\mathbf{c}, \mathbf{p}) \mid \mathbf{p} \in \mathbf{P}_{re}\}$. The spatial probability [**SMITH2014**] for interior and exterior points relative to the centroid are calculated as:

$$\mathbf{S}(\mathbf{P_I} \mid \mathbf{c}) = exp\left(-\frac{1}{|\mathbf{D}_i|} \sum_{d \in \mathbf{D}_i} d\right) \tag{6}$$

$$\mathbf{S}(\mathbf{P_E} \mid \mathbf{c}) = exp\left(-\frac{1}{|\mathbf{D}_e|} \sum_{d \in \mathbf{D}_e} d\right) \tag{7}$$

$$\mathbf{S_i} = \mathbf{S}(\mathbf{P_I} \mid \mathbf{c}) = \frac{\mathbf{S}(\mathbf{P_I} \mid \mathbf{c})}{\mathbf{S}(\mathbf{P_I} \mid \mathbf{c}) + \mathbf{S}(\mathbf{P_E} \mid \mathbf{c})} \tag{8}$$

$$\mathbf{S_e} = \mathbf{S}(\mathbf{P_E} \mid \mathbf{c}) = \frac{\mathbf{S}(\mathbf{P_E} \mid \mathbf{c})}{\mathbf{S}(\mathbf{P_I} \mid \mathbf{c}) + \mathbf{S}(\mathbf{P_E} \mid \mathbf{c})} \tag{9}$$

The interior and exterior score of a point is calculated as:

$$\mathbf{I} = (\mathbf{S_i} \cdot \mathbf{w}) + \frac{|\mathbf{P}_{ri}|}{|\mathbf{P}_{ri}| + |\mathbf{P}_{re}|} \tag{10}$$

$$\mathbf{E} = (\mathbf{S_e} \cdot \mathbf{w}) + \frac{|\mathbf{P}_{ri}|}{|\mathbf{P}_{ri}| + |\mathbf{P}_{re}|} \tag{11}$$

Refinement $(p,\ P_I, P_E, r, w)$ as $\mathbb{R}^2 \times \mathcal{P}(\mathbb{R}^2) \times \mathcal{P}(\mathbb{R}^2) \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^3$

$$= \begin{cases} \text{Refinement}\,(p, P_I, P_E, 1.1r, w)\,, \text{if } F_i = F_e = 0 \\ [1, 0, 1]\,, \text{if } a + b = 1 \wedge a < b \wedge F_i = F_e \wedge \Psi \\ [1, S_i, S_e + 0.3]\,, \text{if } a + b \neq 1 \wedge a < b \wedge F_i = F_e \wedge \Psi \\ [1, 1, 0]\,, \text{if } a + b = 1 \wedge a \not< b \wedge F_i = F_e \wedge \Psi \\ [1, S_i + 0.3, S_e]\,, \text{if } a + b = 1 \wedge a \not< b \wedge F_i = F_e \wedge \Psi \\ [1, 0.5, 0.5]\,, \text{if } I + E < 3 \wedge (S_i = 1 \vee S_e = 1) \wedge p \in P_I \\ [2, 0.5, 0.5]\,, \text{if } I + E < 3 \wedge (S_i = 1 \vee S_e = 1) \wedge p \in P_E \\ [1, S_i, S_e]\,, \text{if } I + E < 3 \wedge ((S_i = S_e) \vee (S_i > S_e)) \wedge p \in P_I \\ [2, S_i, S_e]\,, \text{if } I + E < 3 \wedge ((S_i = S_e) \vee (S_i < S_e)) \wedge p \in P_E \end{cases} \tag{12}$$

$= (e1, e2, e3).e1 \in \{1, 2\}$ and $e2, e3 \in [0, 1]$ and $\Psi = ((F_i < 0.55 \wedge F_e < 0.55) \vee (|P_{ri}| + |P_{re}| < 3))$ and $(a, b) = \mathit{Retrieve}(\mathbf{c}, \mathbf{P_I}, \mathbf{P_E}, r)$

3 Results

The proposed model was evaluated using **sklearn.metrics** for binary classification. The results showed that the proposed model outperforms top existing models from SciKit library, and it's visual representation of the refined subsets shows a better class separation.

Method	TN	FN	TP	FP	ACC	PRE	REC	F1	MCC
PERSIST(proposed)	991.2	60.4	939.6	8.8	0.966	0.992	0.939	0.965	0.933
K-Nearest Neighbors	962.94	49.28	950.72	37.06	0.96	0.96	0.95	0.96	0.91
Random Forest	978.12	76.42	923.58	21.88	0.95	0.98	0.92	0.95	0.9
Decision Tree	967.82	90.88	909.12	32.18	0.94	0.97	0.91	0.94	0.88
Radius Neighbors	988.56	109.98	890.02	11.44	0.94	0.99	0.89	0.94	0.88
Kernel Density	931.72	169.62	830.38	68.28	0.88	0.92	0.83	0.87	0.77
Neural Net MLP	906.22	187.82	812.18	93.78	0.86	0.9	0.81	0.85	0.72

Table 1. Average performance of different models for data denoising

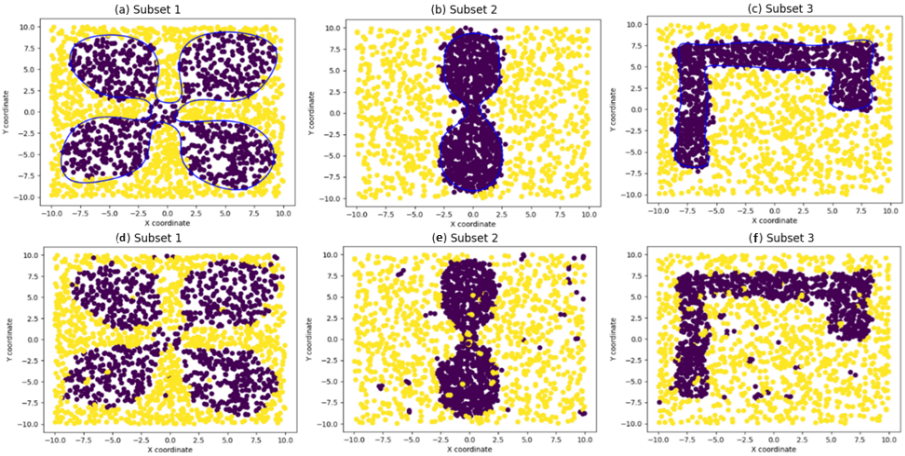


Figure 3. Persist model (a),(b),(c) and K-Nearest Neighbors (d),(e),(f)

Sergiu Scrob

Technical University of Moldova, Chisinau, Moldova
E–mail: scrobsergiu@gmail.com
ORCID: https://orcid.org/0000-0001-6955-0607