Density-Based Fuzzy Clustering as a First Step to Learning Rules: Challenges and Solutions

Gözde Ulutagay[1] and Vladik Kreinovich[2]
[1]Department of Industrial Engineering, Izmir University
Izmir, Turkey, gozde.ulutagay@gmail.com
[2]University of Texas at El Paso
El Paso, Texas 79968, USA, vladik@utep.edu

1. Clustering is How We Humans Make Decisions

- Most algorithms for control and decision making take, as input, the *values* of the input parameters.
- In contrast, we normally only use a *category* to which this value belongs; e.g., when we select a place to eat:
  - instead of exact prices, we consider whether the restaurant is cheap, medium, or expensive;
  - instead of details of food, we check whether it is Mexican, Chinese, etc.
- When we select a hotel, we take into account how many stars it has, is it walking distance from the conf. site.
- First, we *cluster* possible situations, i.e., divide them into a few groups.
- Then, we make a decision based on the group to which the current situation belongs.

2. Clustering is a Natural First Step to Learning the Rules

- Computers process data much faster than we humans.
- However, e.g., in face recognition, we are much better than the best of the known computer programs.
- It is thus reasonable to emulate the way we humans make the corresponding decisions; e.g.:
  - to first cluster possible situations,
  - and then make a decision based on the cluster containing the current situation.

3. Clustering: Ideal Case

- Each known situation is described by the values $x = (x_1, \ldots, x_n)$ of n known quantities.
- When we have many situations, we can talk about the *density* $d(x)$: # situations per unit volume.
- Clusters are separated by voids: there are cats, there are dogs, but there is no continuous transition.
- Within each cluster, we have $d(x) > 0$.
- Outside clusters, we have $d(x) = 0$. So:
  - once we know the density $d(x)$ at each point x,
  - we can find each cluster as the connected component of the set $\{x : d(x) > 0\}$.

4. Clustering: A More Realistic Case

- We often have objects in between clusters.
- For example, coughing and sneezing patients can be classified into cold, allergy, flu, etc.
- However, there are also rare diseases.
- Let t be the density of such rare cases.
  - If $d(x) < t$, then most probably x is not in any major cluster.
  - If $d(x) > t$, then some examples come from one of the clusters that we are trying to form.
- Resulting clustering algorithm:
  - we select a threshold t, and
  - we find each cluster as a connected component of the set $\{x : d(x) \geq t\}$.

5. How to Estimate the Density $d(x)$

- In practice, we only have finitely many examples $x^{(1)}, \ldots, x^{(N)}$.
- The measured values $x^{(j)}$ are, in general, different from the actual (unknown) values x.
- Let $\rho(\Delta x)$ be the probability density of meas. errors.
- Then, for each j, the probability density of actual values is $\rho(x^{(j)} - x)$.
- Observations are equally probable, so

$$d(x) = p(x^{(1)}) \cdot \rho(x^{(1)} - x) + \ldots + p(x^{(N)}) \cdot \rho(x^{(N)} - x) = \frac{1}{N} \cdot \sum_{j=1}^{N} \rho(x^{(j)} - x).$$

- This formula is known as the *Parzen window*.
- The corresponding function $\rho(x)$ is known as a *kernel*.

6. Resulting Clustering Algorithm

- At first, we select a function $\rho(x)$.
- Then, based on the observed examples $x^{(1)}$, $x^{(2)}$, $\ldots$, $x^{(N)}$, we form a density function

$$d(x) = \frac{1}{N} \cdot \sum_{j=1}^{N} \rho(x^{(j)} - x).$$

- After that, we select a threshold t.
- We find the clusters as the connected components of the set $\{x : d(x) \geq t\}$.
- For imprecise ("fuzzy") expert estimates, instead of probabilities, we have membership functions.
- So, we get similar formulas.

7. Discussion

- Empirical results:
  - The best kernel is the Gaussian function $\rho(x) \sim \exp(-\text{const} \cdot x^2)$.
  - The best threshold t is the one for which clustering is the most robust to selecting t.
- Our 1st challenge is to provide a theoretical explanation for these empirical results.
- 2nd challenge: take into account that some observations may be erroneous.
- 3rd challenge: clustering algorithms should return *degrees* of belonging to different clusters.
- 4th challenge: hierarchy – animals should be first classified into dangerous and harmless, then further.

8. Why Gaussian Kernel: A Solution to the 1st Part of the 1st Challenge

- The Gaussian distribution of the measurement error is indeed frequently occurring in practice.
- This empirical fact has a known explanation:
  - a measurement error usually consists of a large number of small independent components, and,
  - according to the Central Limit theorem:
    - the distribution of the sum of a large number of small independent components
    - is close to Gaussian.
- Expert inaccuracy is also caused by a large number of relatively small independent factors.

9. Alternative Explanation

- We start with the discrete empirical distribution $d_N(x)$ in which we get N values $x^{(j)}$ with equal probability.
- We "smoothen" $d_N(x)$ by convolving it with the kernel function $\rho(x)$: $d(x) = \int d_N(y) \cdot \rho(x - y)\, dy$.
- This works if we properly select the half-width σ of the kernel:
  - if we select a very narrow half-width, then each original point $x^{(j)}$ becomes its own cluster;
  - if we select a very wide half-width, then we end up with a single cluster.
- The choice of this half-width is usually performed empirically:
  - we start with a small value of half-width, and
  - we gradually increase it.

10. Alternative Explanation (cont-d)

- Since the kernel functions are close to each other, the resulting convolutions are also close.
- So, it is computationally efficient to apply a small modifying convolution to the previous convolution result.
- The resulting convolution is the result of applying a large number of minor convolutions.
- From the mathematical viewpoint, a convolution means adding an independent random variable.
- Applying a large number of convolutions is equivalent to adding many small random variables.
- Thus, it is equivalent to adding Gaussian variable – i.e., to Gaussian convolution.

11. Explaining the Empirical Formula for the Best Threshold

- Empirically, the best threshold t is the one for which the largest change in t keeps clusters intact.
- This change is equal to the difference between two neighboring values $d(x)$.
- We have $d(x)$ values per unit volume, so the next one is at distance $d(x)^{-1/n}$.
- The difference between the neighboring values $d(x)$ is thus proportional to $\|\nabla d(x)\| \cdot d(x)^{-1/n}$.
- So, we select $t = \max\limits_x \|\nabla d(x)\| \cdot d(x)^{-1/n}$.
- In general, we could consider $t = \max\limits_x f(\nabla d(x), d(x))$.
- Rotation- and scale-invariance imply $f(a, b) = \|a\|^\alpha \cdot b^\beta$.
- This explains the empirically formula.

12. What If Not All Objects Are of Given Type

- In the above algorithm, we assumed that each measurement represents the situation of the given type.
- In practice, we are not always sure that what we measured in necessarily one of such situations.
- For each each observed situation j, we can estimate the probability p_j that this situation is of given type.
- It is desirable to take these probabilities into account during clustering.
- In Parzen's formula, we selected each point $x^{(j)}$ with equal probability $p(x^{(j)}) = \dfrac{1}{N}$.
- *Idea:* select probability $p(x^{(j)}) \sim p_j$; then, $d_0(x) \stackrel{\text{def}}{=} \sum\limits_{j=1}^{N} p_j \cdot \rho(x^{(j)} - x)$, where $k = 1/\left(\sum\limits_{j=1}^{N} p_j\right)$.

13. Towards Fuzzy Clusters

- *In the above algorithm:* a cluster c is determined as a connected component of the set $\{x : d(x) \geq t\}$.
- *Observation:* from the finite sample, we can determine $d(x)$ and t only approximately.
- *Conclusion:* if we can connect $x \notin c$ with c by a path with $d(x) \geq t - \varepsilon$, it is probable that $x \in c$.
- *Idea:* take the ratio $\dfrac{t - \varepsilon}{t}$ as the degree $d_c(x)$ to which x is in the cluster c.
- *In precise terms:* as $d_c(x)$, we take the max of ratios $\dfrac{s}{t}$ over all paths with $d(x) \geq s$ that connect x to c.
- *Equivalent definition:* s is the largest for which both x and c belong to the same connected component of

$$\{x : d(x) \geq s\}.$$

14. Towards Hierarchical Clustering

- *Idea:* it is desirable to come up with sub-clusters of each cluster c.
- *Possible solution:* use larger thresholds $t' > t$, and find connected components of the set $\{x : d(x) \geq t'\}$.
- *Fact:* most of these ideas have been applied to real-life problems.
- *Result:* the resulting clustering is indeed
  - closer to the expert-generated clustering
  - than the clustering performed by the usual fuzzy clustering algorithms.

15. Acknowledgments

- This work was supported in part:
  - by NSF grants HRD-0734825 and HRD-1242122,
  - by Grants 1 T36 GM078000-01 and 1R43TR000173-01 from NIH,
  - by a grant N62909-12-1-7039 from ONR, and
  - by a grant 111T273 from TUBITAK.
- This work was partially performed when Gözde Ulutagay was visiting the University of Texas at El Paso.
- The authors are greatly thankful to Professor Zadeh for inspiring discussions.