Attentive Clustering Processes: Part 2

Dominic Danks[1,2]

[1]The Alan Turing Institute

[2]University of Birmingham

November 13, 2020

My role as introducer 2

- To "get the discussion going"...

My role as introducer 2

- To "get the discussion going"...

- Two-part discussion: 1) Clarification/understanding; 2) Impressions/opinion

My role as introducer 2

- To "get the discussion going"...
- Two-part discussion: 1) Clarification/understanding; 2) Impressions/opinion
- Facilitating 1):
  - Give my take on understanding the key points and highlight aspects of the paper(s) which I think are least clear or could cause confusion.

My role as introducer 2

- To "get the discussion going"...
- Two-part discussion: 1) Clarification/understanding; 2) Impressions/opinion
- Facilitating 1):
  - Give my take on understanding the key points and highlight aspects of the paper(s) which I think are least clear or could cause confusion.
- Facilitating part 2):
  - Highlight aspects of the paper(s) which I think are likely to draw strong and/or differing opinions depending on taste and primary interests.

High-level reading

- This paper is part of a small series of recent papers considering the problem of *amortized clustering* which uses *labelled* training data.

High-level reading

- This paper is part of a small series of recent papers considering the problem of *amortized clustering* which uses *labelled* training data.
  - Q: How do "amortized clustering" and "classification" compare/differ?

High-level reading

- This paper is part of a small series of recent papers considering the problem of *amortized clustering* which uses *labelled* training data.
  - Q: How do "amortized clustering" and "classification" compare/differ? Amortized clustering can be applied to *test* data *without* known categories to potentially uncover the "natural" categories.

High-level reading

- This paper is part of a small series of recent papers considering the problem of *amortized clustering* which uses *labelled* training data.
  - Q: How do "amortized clustering" and "classification" compare/differ? Amortized clustering can be applied to *test* data *without* known categories to potentially uncover the "natural" categories.

- It builds on the "Neural Clustering Processes" paper's Clusterwise Clustering Process (CCP) by proposing a more expressive encoding of the point assignment when considering cluster k.

High-level reading

- More specifically, it proposes not using the quantities

$$U = \sum_{i=1}^{m_k} u\left(x_{a_i}\right) \text{ and } G = \sum_{j=1}^{k-1} g\left(\sum_{i \in \mathrm{s}_j} h\left(x_i\right)\right)$$

as input to NNs approximating $p_\theta\left(\mathrm{z}_k \mid \mathrm{x}_k\right)$ and $p_{\theta,i}\left(b_i \mid \mathrm{z}_k, \mathrm{x}_k\right)$, but rather to define

$$\left(\bar{u}_{d_k}, \bar{u}_1 \ldots \bar{u}_{m_k}\right) = \mathrm{ISAB}\left[u\left(x_{d_k}\right), u\left(x_1\right) \ldots u\left(x_{m_k}\right)\right]$$

$$\bar{\mathrm{u}}_{\mathrm{a}} = \left(\bar{u}_1 \ldots \bar{u}_{m_k}\right)$$

and use

$$U = \mathrm{PMA}\left(\mathrm{MAB}\left(\bar{\mathrm{u}}_{\mathrm{a}}, \bar{u}_d\right)\right), \quad G = \sum_{j=1}^{k-1} g\left(\mathrm{PMA}\left(h\left(\bar{x}_i\right), i \in \mathrm{s}_j\right)\right).$$

High-level reading

- More specifically, it proposes not using the quantities

$$U = \sum_{i=1}^{m_k} u\left(x_{a_i}\right) \text{ and } G = \sum_{j=1}^{k-1} g\left(\sum_{i \in \mathrm{s}_j} h\left(x_i\right)\right)$$

as input to NNs approximating $p_\theta\left(\mathrm{z}_k \mid \mathrm{x}_k\right)$ and $p_{\theta,i}\left(b_i \mid \mathrm{z}_k, \mathrm{x}_k\right)$, but rather to define

$$\left(\bar{u}_{d_k}, \bar{u}_1 \ldots \bar{u}_{m_k}\right) = \mathrm{ISAB}\left[u\left(x_{d_k}\right), u\left(x_1\right) \ldots u\left(x_{m_k}\right)\right]$$

$$\bar{\mathrm{u}}_{\mathrm{a}} = \left(\bar{u}_1 \ldots \bar{u}_{m_k}\right)$$

and use

$$U = \mathrm{PMA}\left(\mathrm{MAB}\left(\bar{\mathrm{u}}_{\mathrm{a}}, \bar{u}_d\right)\right), \quad G = \sum_{j=1}^{k-1} g\left(\mathrm{PMA}\left(h\left(\bar{x}_i\right), i \in \mathrm{s}_j\right)\right).$$

Q: The motivation for this is that the original sum for U should not uniformly weight available points, but should focus on those most similar to anchor x_{d_k}. Does that intuition clearly propagate through these complicated attention mechanisms?

High-level reading

- I found the details around how to *train* the model, i.e. what sorts of datasets were required, how to generate them etc. very sparse. Perhaps there is more information elsewhere? To me this is an extremely important aspect which is not given a great deal of attention.

High-level reading

- I found the details around how to *train* the model, i.e. what sorts of datasets were required, how to generate them etc. very sparse. Perhaps there is more information elsewhere? To me this is an extremely important aspect which is not given a great deal of attention. Q: Do you agree?

High-level reading

- I found the details around how to *train* the model, i.e. what sorts of datasets were required, how to generate them etc. very sparse. Perhaps there is more information elsewhere? To me this is an extremely important aspect which is not given a great deal of attention. Q: Do you agree?

- Q: How do your main takeaways compare to mine?

NCPs vs CCPs vs ACPs

How do Neural Clustering Processes (NCPs), Clusterwise Clustering Processes (CCPs) and Attentive Clustering Processes (ACPs) compare/contrast?

NCPs vs CCPs vs ACPs

How do Neural Clustering Processes (NCPs), Clusterwise Clustering Processes (CCPs) and Attentive Clustering Processes (ACPs) compare/contrast?

- CCPs vs ACPs is covered in the paper and in our discussions so far — NCPs not so much. They are dismissed due to their $O(N)$ scaling.

NCPs vs CCPs vs ACPs

How do Neural Clustering Processes (NCPs), Clusterwise Clustering Processes (CCPs) and Attentive Clustering Processes (ACPs) compare/contrast?

- CCPs vs ACPs is covered in the paper and in our discussions so far — NCPs not so much. They are dismissed due to their $O(N)$ scaling.
  - Q: NCPs may scale as O(N), but how terminal is this for the applications we would consider? What about a wall-clock time for the sorts of data we would deal with, including training?

NCPs vs CCPs vs ACPs

How do Neural Clustering Processes (NCPs), Clusterwise Clustering Processes (CCPs) and Attentive Clustering Processes (ACPs) compare/contrast?

- CCPs vs ACPs is covered in the paper and in our discussions so far — NCPs not so much. They are dismissed due to their $O(N)$ scaling.
  - Q: NCPs may scale as O(N), but how terminal is this for the applications we would consider? What about a wall-clock time for the sorts of data we would deal with, including training?

- NCPs do not require i) the use of De Finetti's theorem to posit a posterior approximation or ii) an additional latent variable z over which a posterior must be estimated. They also explicitly provide a posterior probability, whereas with CCPs/ACPs this must be post-hoc estimated. Note also that NCPs tend to perform very comparably to ACPs and better than CCPs.

NCPs vs CCPs vs ACPs

How do Neural Clustering Processes (NCPs), Clusterwise Clustering Processes (CCPs) and Attentive Clustering Processes (ACPs) compare/contrast?

- CCPs vs ACPs is covered in the paper and in our discussions so far — NCPs not so much. They are dismissed due to their $O(N)$ scaling.
  - Q: NCPs may scale as O(N), but how terminal is this for the applications we would consider? What about a wall-clock time for the sorts of data we would deal with, including training?

- NCPs do not require i) the use of De Finetti's theorem to posit a posterior approximation or ii) an additional latent variable z over which a posterior must be estimated. They also explicitly provide a posterior probability, whereas with CCPs/ACPs this must be post-hoc estimated. Note also that NCPs tend to perform very comparably to ACPs and better than CCPs. Q: Does this make NCPs more natural/elegant than CCPs/ACPs?

Overall thoughts

- I like the idea of meta-learning in the context of clustering. However, given that it is a relatively radical shift in approach to clustering I would like to see the papers dwell more upon the training procedure and data generation.

Overall thoughts

- I like the idea of meta-learning in the context of clustering. However, given that it is a relatively radical shift in approach to clustering I would like to see the papers dwell more upon the training procedure and data generation. Q: Perhaps they have elsewhere or this is not as new as I think?

Overall thoughts

- I like the idea of meta-learning in the context of clustering. However, given that it is a relatively radical shift in approach to clustering I would like to see the papers dwell more upon the training procedure and data generation. Q: Perhaps they have elsewhere or this is not as new as I think?

- I find NCPs more elegant than CCPs and ACPs and am interested in seeing more about how prohibitive NCPs are to train and apply in practice.

Overall thoughts

- I like the idea of meta-learning in the context of clustering. However, given that it is a relatively radical shift in approach to clustering I would like to see the papers dwell more upon the training procedure and data generation. Q: Perhaps they have elsewhere or this is not as new as I think?

- I find NCPs more elegant than CCPs and ACPs and am interested in seeing more about how prohibitive NCPs are to train and apply in practice. Q: Does anyone know more about this?

Overall thoughts

- I like the idea of meta-learning in the context of clustering. However, given that it is a relatively radical shift in approach to clustering I would like to see the papers dwell more upon the training procedure and data generation. Q: Perhaps they have elsewhere or this is not as new as I think?
- I find NCPs more elegant than CCPs and ACPs and am interested in seeing more about how prohibitive NCPs are to train and apply in practice. Q: Does anyone know more about this?
- Whilst the above still holds, transformer-based approaches are seeing great success in other settings (e.g. GPTs), so the development of ACPs is understandable.

Discussion

Some of the previously posed content to start us off:

Discussion

Some of the previously posed content to start us off:

- Overall thoughts on the paper(s)
- Amortized clustering as a concept — do you like it/believe it can generalise sufficiently?
- Is the attention aspect of ACPs an intuitive, clever tweak on CCPs or a significant, uninterpretable complication?
- Is there sufficient discussion of what is required to train an amortized clustering algorithm, e.g. in terms of data generation, style, generality, compute requirements?
- NCPs vs CCPs vs ACPs: which do you prefer and why?
- Other comments, thoughts, questions, opinions?