Explainable Unsupervised Learning as Constraint Satisfaction

Ian Davidson[1] **and S. S. Ravi**[2]

[1]Department of Computer Science, University of California Davis, Davis, CA 95616.
[2]Biocomplexity Institute and Initiative, University of Virginia, Charlottesville, VA 22904
and Department of Computer Science, University at Albany – SUNY, Albany, NY 12222.
davidson@cs.ucdavis.edu, ssravi0@gmail.com

Abstract

The need for explanation is paramount in machine learning as machines aid and sometimes even replace humans in decision making. Our recent papers on this topic for global level explanation explore two forms of XAI: (i) Explanation via post-processing in NeurIPS 2018 (Davidson, Gourru, and Ravi 2018) and AAAI 2020 (Sambaturu et al. 2020) and (ii) Explanation by design in IJCAI 2018 (Dao et al. 2018) and IJCAI 2021 (Zhang and Davidson 2021). In this work we try to explain a group/cluster by discovering/learning a constraint that is satisfied by each group; thus, the constraint serves as the explanation. We overview each style of explanation and present several challenges we hope that the CP community can help with. Note though this work is predominantly in the area of clustering, the results can be applied to any ML algorithm that produces a k-block set partition as its output.

1 Introduction and Previous Work

As AI permeates society and is used in more complex settings, the need for explanation becomes paramount. The area of XAI for supervised learning has been well studied and focuses on several well known questions shown in the first column of Table 1. This work aims to explain a model at the local/instance level by answering questions such as "Why is this particular image classified as a frog?". Our recent work has explored these XAI questions from the perspective of global explanations by trying to explain a class ("What is a frog?") and what differentiates classes from each other. The focus on global level explanation lends itself to constraint satisfaction style problems of the following form: does there exist a subset of tags/descriptors that covers all instances in a group?

We begin by summarizing previous work and then move onto our own work. Finally, we sketch potential future directions that we hope the CP community can contribute to.

Previous Work –2010 and Earlier. Figure 1 shows a pictorial overview of the area. Explanation by design algorithms were some of the earliest problems studied in machine learning under the title *conceptual grouping* (Michalski and Stepp 1983; Fisher 1987). This work tried to simultaneously find a grouping and a description but was limited to categorical data. Initial work led to heuristic formulations and did not scale well; however, the resultant explanation could sometimes be interpreted as an ontology. The area was restarted later (Mueller and Kramer 2010; Ouali et al. 2016) with a focus on optimization. This work formulated integer linear programs (ILPs) using the results of frequent pattern of mining to pre-process the data so as to represent each instance as a collection of concepts. The optimization focus was to choose the fewest number of concepts to form a group, with all the instances in a group having the concepts associated with the group. The constraints enabled coverage style requirements, namely that an instance belongs precisely to one group. But again all this work required the features to be human interpretable.

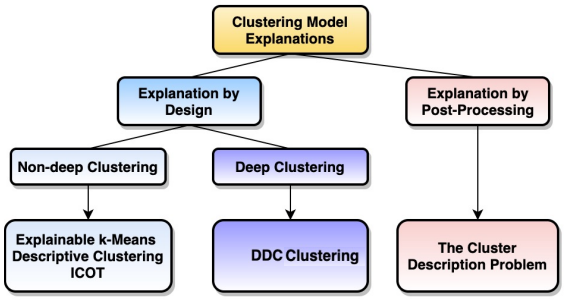

Figure 1: A high level overview of the work on grouping and explanation.

2 Work in the Last Four Years

Explanation by design algorithms are relatively well studied. The work of (Laber and Martinho 2020; Dasgupta et al. 2020) show how to simultaneously build a k-means grouping and a decision tree with strong guarantees. Whilst a clever heuristic method to simultaneously group the data and maximize a form of interpretability is used in (Saisubramanian, Galhotra, and Zilberstein 2020), this work also requires features to be interpretable.

Our recent papers at Neurips 2018/AAAI 2020 (David-

XAI Question	Analogous Question in Unsupervised Learning
Why did you do that?	Explain why you formed these groups?
Why not something else?	Why did you not place these instances into a group? Can you make this group explanation simpler? Can you generate another group explanation?
When do you succeed/fail? When can I trust you?	Which groups and/or points are not well explained? Are these explanations robust/stable?

Table 1: The analogous questions in the unsupervised learning setting for XAI supervised learning questions (Gunning 2017).

son, Gourru, and Ravi 2018; Sambaturu et al. 2020) and IJCAI 2018/2021 (Dao et al. 2018) (Zhang and Davidson 2021) explore a different form of explanation where unsupervised learning is performed on one set of features and explanation is done on another set referred to as tags. This is suitable for a variety of settings when the features generated are private or not interpretable (e.g., from a deep embedding). The requirement of tags can be fulfilled via automatic generation or hand annotation; for example, images can be auto-captioned and graphs can have vertex labels. Our work tries to discover constraints which are associated with one group and not others; these constraints are then considered as the explanation for the group.

2.1 Explanation by Post-Processing

The area of post-processing to obtain explanations has been relatively understudied for unsupervised learning. We first formulated the idea of explanation as a set coverage requirement/constraint in NeurIPS 2018 (Davidson, Gourru, and Ravi 2018) and later developed efficient approximation algorithms (Sambaturu et al. 2020) as well as applications to social networks (Davidson, Gourru, and Velcin 2020).

Consider the example shown in Figure 2 where Twitter accounts are partitioned (say based on the follower relationship) into two groups (red and blue) and are to be explained using their hashtag usage. The edges show the hashtag usage of each account. Informally, we wish to select a subset Y_1 of yellow tags so that each of the red instances uses at least one tag from Y_1 (i.e., the yellow tags in Y_1 cover the red items) and a *different* subset Y_2 of yellow tags (disjoint from Y_1) that cover the blue instances. The disjointness requirement is important so that the explanation is not only for what is in the group but also for what makes the groups different.

We formalize this as a set cover style problem in (Davidson, Gourru, and Ravi 2018) which we refer to as the Disjoint Tag Descriptor Feasibility (DTDF) problem.

Result #1: The DTDF Problem is Computationally Intractable. We have shown (Davidson, Gourru, and Ravi 2018) that the DTDF problem is **NP**-hard for even simple cases as indicated in the following theorem. However, a simple ILP formulation exists.

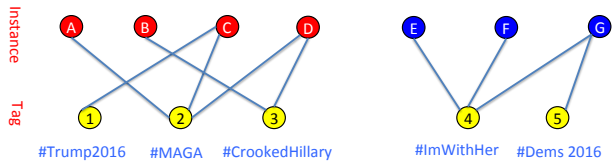

Figure 2: A simple Twitter network example with two (red/blue) groups/communities to be explained using the hashtags of each individual user. For example, person A uses `MAGA`, person B uses `CrookedHillary` and so on. A valid explanation for the red community is {`MAGA, CrookedHillary`} and for the blue community is {`ImWithHer`}.

Theorem 2.1 ((Davidson, Gourru, and Ravi 2018)) *The* DTDF *problem is **NP**-complete even when the number of groups is 2 and the tag set of each item has at most 3 tags.*

2.2 Explanation by Design

Here we explore work on simultaneous grouping and explanation. This work is particularly important as a very good grouping need not have any reasonable explanation. Our initial work (Dao et al. 2018) on this topic uses the setting of finding a grouping on a set of features (which can be a deep learning embedding) and explaining it using an auxiliary set of tags. We viewed this as a Pareto optimization problem with one objective to find a compact grouping and the other to find a compact explanation. Our subsequent work (Zhang and Davidson 2021) looked at this setting using deep learning for both grouping and explanations by generating only one solution.

As stated above, the goal of descriptive grouping is to find a good grouping along with an explanation. The grouping method uses the features of instances while the resulting explanation uses auxiliary tags associated with instances. Naturally, the problem formulation involves two objectives, one to measure the quality of the grouping and the other to quantify the effectiveness of the explanation. Our first work (Dao et al. 2018) generates the Pareto frontier corresponding to the two objectives and allows a user to pick the appropriate point in the frontier. In our second work (Zhang and Davidson 2021) we study a similar setting as before, but using a deep learning formulation. Since we do not try to find the Pareto front, this work scales to large data sets.

Result #2: We develop an ILP formulation and a significantly more efficient constraint programming (CP) formulation for the bi-objective descriptive grouping problem. The CP formulation uses global constraints to make the search with respect to the explanation objective more efficient.

Result #3: We develop an iterative scheme to compute the complete Pareto frontier with respect to the two objectives. This scheme takes into account the fact that improvements in the grouping objective may be very small (since the objective takes on floating point values).

3 Conclusion and Challenges for the CP Community

XAI is an important emerging area in AI and very understudied in unsupervised learning. The work of ourselves and others have made solid progress but there are still key open questions.

Our work explains a clustering using a set of tags by finding a constraint/explanation that is satisfied for each group/cluster. This context gives rise to two main challenges. The first of these is computational and the second is in terms of complexity of constraints/explanations. The first challenge is to explore whether more efficient CP formulations can be developed for our ILP formulations. The second, that is, the explanatory language challenge, is more complex. One can view our existing work as discovering a disjunction and future challenges will explore more complex languages/constraints as the following:

- Adding meta-information to the description such as significance/weight and type. This will then allow us to include additional constraints on the explanation such as requiring the total significance to exceed a minimum threshold and to ensure diversity based on type.
- Developing explanations beyond simple disjunctions into conjunctive normal form (i.e., a conjunction of disjunctions). This will allow more complex descriptions.
- Extending explanations to be beyond simple tag style information to evolving tags (over time) and continuous and even relational data. For example, for our evaluation on Twitter data we can explore frequency of tag usage and co-occurrence of tags.
- Explanations based on other artifacts beyond a single cluster such as notions of core instances, pairs of clusters and overlap between clusters. This can be used to explain complicated clusterings to a general audience.

Finally, perhaps the greatest challenge is not to just generate an explanation, but rather to generate explanation that is trustworthy. How to measure trust here is of course a very nebulous challenge and we see several core directions including measuring the stability of the explanation (intrinsic trust) and allowing a human to query the explanation with say counter-factual queries (extrinsic trust).

Acknowledgment: This work is supported in part by NSF Grants IIS-1908530 and IIS-1910306 titled "Explaining Unsupervised Learning: Combinatorial Optimization Formulations, Methods and Applications".

References

Dao, T.; Kuo, C.-T.; Ravi, S. S.; Vrain, C.; and Davidson, I. 2018. Descriptive Clustering: ILP and CP Formulations with Applications. In *Proc. IJCAI*, 1263–1269.

Dasgupta, S.; Frost, N.; Moshkovitz, M.; and Rashtchian, C. 2020. Explainable k-means and k-medians clustering. In *Proc. 37th ICML*, 12–18.

Davidson, I.; Gourru, A.; and Ravi, S. S. 2018. The Cluster Description Problem - Complexity Results, Formulations and Approximations. In *Proc. NeurIPS*, 6193–6203.

Davidson, I.; Gourru, A.; and Velcin, J. 2020. Behavioral Differences - Insights, Explanations and Comparisons of French and US Twitter Usage During Elections. In *J. Social Network Analysis and Mining (to appear)*.

Fisher, D. H. 1987. Knowledge acquisition via incremental conceptual clustering. *Machine learning*, 2(2): 139–172.

Gunning, D. 2017. Explainable Artificial Intelligence (XAI). DARPA Program Update Document. Available from https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.

Laber, E.; and Martinho, L. 2020. On the price of explainability for some clustering problems. In *Proc. 38th ICML*, 5915–5925.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 951–958. IEEE.

Michalski, R. S.; and Stepp, R. 1983. Learning from Observation: Conceptual Clustering. In Michalski, R. S.; Carbonell, J.; and Mitchell, T., eds., *Machine Learning: An Artificial Intelligence Approach*, 331–363. Palo Alto, CA: Tioga Publishing Co.

Mueller, M.; and Kramer, S. 2010. Integer Linear Programming Models for Constrained Clustering. In *DS 2010*, 159–173.

Ouali, A.; Loudni, S.; Lebbah, Y.; Boizumault, P.; Zimmermann, A.; and Loukil, L. 2016. Efficiently finding conceptual clustering models with integer linear programming. In *IJCAI 2016*, 647–654.

Saisubramanian, S.; Galhotra, S.; and Zilberstein, S. 2020. Balancing the Tradeoff Between Clustering Value and Interpretability. In *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, 351–357.

Sambaturu, P.; Gupta, A.; Davidson, I.; Ravi, S. S.; Vullikanti, A. K.; and Warren, A. 2020. Efficient Algorithms for Generating Provably Near-Optimal Cluster Descriptors for Explainability. In *Proc. AAAI*, 1636–1643.

Wang, X.; Qian, B.; Ye, J.; and Davidson, I. 2013. Multi-objective multi-view spectral clustering via Pareto optimization. In *ICDM*, 234–242.

Zhang, H.; and Davidson, I. 2021. Deep Descriptive Clustering. In *Proc. IJCAI*, 3342–3348. ijcai.org.