Hierarchical Data Analysis Using Discernibility Relation

Raymond So

School of Science and Technology, The Open University of Hong Kong, Hong Kong, China

ABSTRACT

Human-in-the-Loop (HITL) machine learning uses human feedback to improve performance of machine learning models. One of the focuses in HITL machine learning research is to explore ways to capture human feedback and transform feedback to useful information that can inform the learning process. This paper outlines a clustering method that is based on discernibility relation in rough set theory. This clustering method presents intermediate clustering results using indiscernibility definition graph. Human users can provide feedback by manipulating the cluster representatives that are presented in an indiscernibility definition graph. Discernibility relation offers a more intuitive understanding of clustering results when compared to distance-based relationship in terms of providing useful feedback to inform the clustering algorithm about its performance.

Keywords: *Human-in-the-Loop machine learning, Hierarchical clustering, Discernibility, Rough Sets.*

I. INTRODUCTION

There are often multiple ways to cluster a dataset. Many believe that clustering is, by nature, subjective undertaking. Most popular clustering algorithms are unsupervised. Domain or application specific information,

which is useful to produce the most desirable clustering results, is sometimes difficult to be integrated into the clustering process. Human-in-the-Loop (HITL) machine learning research aims at tackling this problem. HITL machine learning allows humans to interact with machine learning algorithms with the objective to elicit useful human feedback that can be used to produce the most satisfying results.

One of the challenges in HITL machine learning is interaction design. This involves finding the most intuitive way to present the intermediate machine learning models, elicit human feedback, and transform feedback into the appropriate parametric requirements that can be used to improve the final model. Issues in HITL clustering become even more challenging when a large number of binary attributes are involved. Most popular clustering machine learning algorithms use distance, which does not work well for binary attributes, to create clusters. Typical clustering concepts such as centroid, inter- and intra-cluster distances can be difficult to understand for people who have little machine learning knowledge. It would a challenge to ask users for feedback when the quality of solutions is presented using formal and technical terminologies.

This paper outlines a hierarchical clustering algorithm that uses discernibility to cluster datasets that have only binary attributes. This algorithm includes a feedback mechanism which elicits human feedback on intermediate clustering results. The feedback mechanism presents the cluster representatives using an indiscernibility definition graph at the end of each iteration of the algorithm. Human users will be able to modify the cluster membership by interacting with the indiscernibility definition graph.

II. PRIOR WORK

Observation-Level Interaction (OLI) [5] is one of the major types of HITL machine learning. There are two types of OLI: exploratory and expressive interactions [1]. Exploratory interaction allows users to manipulate data objects in their respective clusters. In other words, exploratory interaction does not cause any changes to cluster membership. Expressive interactions, on the other hand, allow users to move data objects into and out of clusters, causing changes to cluster membership. Expressive interaction represents what the user like (or not) to see in cluster membership produced by clustering algorithms. This type of interaction reflects the user's non-parametric requirements (preferences) for clustering. Positive results of an experimental use of expressive interaction in AGNES were reported in [3].

Work reported in [2] suggests that *examples* are most preferred by human users when it comes to understanding machine learning models. Bayesian Case Model (BCM) was used to identify the cluster representatives in a clustering exercise of recipes. The result was verified by human users and reported as satisfactory.

III. DEFINITIONS

Following the definitions in [6], we define data objects S as:

$$S = (U, At, \{V_a \mid a \in At\}, \{I_a \mid a \in At\})$$

where U is a finite set of data objects, At is a finite nonempty set of attributes, V_a is a nonempty set of values for an attribute $a \in At$, and $Ia : U \rightarrow V_a$ is an information function, such that for a data object $x \in U$, an attribute $a \in At$, and a value $v \in V_a$, $I_a(x) = v$ denotes data object x has an attribute a which value is v.

Indiscernibility, according to [Zhao 2007], is defined as:

$$IND(A) = \{(x, y) \in U \times U \mid \forall a \in A, I_a(x) = I_a(y)\}$$

The degree of discernibility r is defined as the cardinality of A, such that $r = |A|$ [4]. We use r to quantify indiscernibility relations.

Table 1 An Example Dataset (U). Each data object has six binary attributes (A).

	a	**b**	**c**	**d**	**e**	**f**
o_1	1	1	1	1	1	1
o_2	1	0	1	0	1	1
o_3	0	0	1	1	0	0
o_4	1	1	1	0	0	1
o_5	1	0	1	0	1	1
o_6	0	0	0	1	1	0
o_7	1	0	1	1	1	1
o_8	0	0	0	0	1	1
o_9	1	0	0	1	0	0

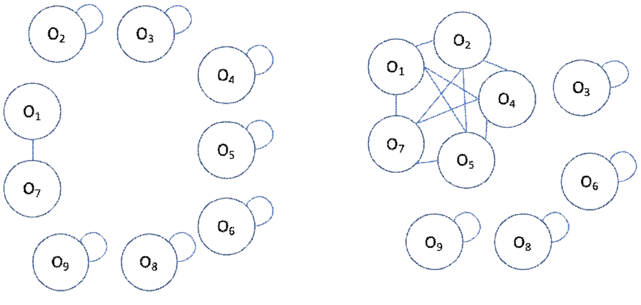
Consider the example data set $U = \{o_1, \dots, o_8\}$ in Table 1. Each data object has six binary attributes $At = \{a, b, c, d, e, f\}$, $V_a = \{1, 0\}$. Data objects o_1 and o_7 in are indiscernible when $r = 5$.

$$IND_{r=5}(\{a, c, d, e, f\}) = \{o_1, o_7\}$$

Indiscernibility relationships can be ranked (or compared). For instance, considering the following indiscernibility relationship:

$$IND_{r=4}(\{a, c, e, f\}) = \{o_1, o_2, o_4, o_5, o_7\}$$

We say $IND_{r=5}$ is *stronger* than $IND_{r=4}$ as data objects in $IND_{r=5}$ have more attribute values in common than in $IND_{r=4}$.



$IND_{r=5}(\{a, c, d, e, f\}) = \{o_1, o_7\}$ $IND_{r=4}(\{a, c, e, f\}) = \{o_1, o_2, o_4, o_5, o_7\}$

Figure 1Indiscernibility Definition Graphs

Indiscernibility relation can be presented using an indiscernibility definition graph (IDG), which is defined in [4] as graph with the elements of V_a as nodes or vertices, and a set of edges $E_a \subseteq V_{a^2}$ such that:

$$IDG_a = (V_a, E_a)$$

An edge $(v_1, v_2) \in E_a$ is to be interpreted as that an object with value v_1 is indiscernible with object with value v_2. The edge between o_1, o_7 in Fig. 1 represents that o_1, o_7 are indiscernible on attributes a, b, d, e, f.

We define *indiscernible clusters* $C_{IND_{r=i}}$ as the collection of indiscernible clusters when $r = i$. The indiscernible clusters in Fig. 1 are:

$$C_{IND_{r=5}} = \{\{o_1, o_7\}, \{o_2\}, \{o_3\}, \{o_4\}, \{o_5\}\{o_6\}, \{o_8\}, \{o_9\}\}$$

$$C_{IND_{r=4}} = \{\{o_1, o_2, o_4, o_5, o_7\}, \{o_3\}, \{o_6\}, \{o_8\}, \{o_9\}\}$$

The *cluster representatives* of $C_{IND_{r=4}}$ are o_1, o_3, o_6, o_8, o_9.

IV. PROPOSED SOLUTIONS

The proposed algorithm begins with placing data objects having the maximum number of indiscernible features $IND_{r=|At|}$ into indiscernible cluster $C_{IND_{r=|At|}}$. Clusters are created for each of the remaining data objects. An indiscernibility graph is used to present the clusters to human users for comments and elicit any expressive feedback. For instance, the human users *may* drag any data objects o_3, o_6, o_8, o_9 in Fig. 1 into the indiscernible cluster $IND_{r=4}(\{a, c, e, f\})$ if they see them as member of the indiscernible cluster. When this happens, the algorithm explores the possibility of accommodating the change of cluster membership by searching into indiscernible clusters with a smaller number of r. The whole algorithm is outlined below.

```
Initialization: $rt \leftarrow |At|$

1. Create indiscernible clusters $C_{IND_{r=rt}}$
2. Create and present the indiscernibility
      graph for $C_{IND_{r=rt}}$
3. $C_k, o_k \leftarrow Feedback\,(C_{IND_{r=rt}})$
4. if $C_k \neq null$ and $o_k \neq null$, then
      while $rt > 1$ do
         $rt \leftarrow rt - 1$
         Create indiscernible clusters for
         $C_{IND_{r=rt}}$
         if $C_k \cup o_k \in C_{IND_{r=rt}}$, then
            Go to Step 1
         else if $rt > 1$, then
            continue
         else
            abort
         end
      end
   else
      $rt \leftarrow rt - 1$
      Go to Step 1
   end
```

V. EXPERIMENT RESULTS

We used the algorithm to create indiscernible clusters of recipes and identify the most representative ingredients of each cluster. We compare our findings with the work reported in [2].

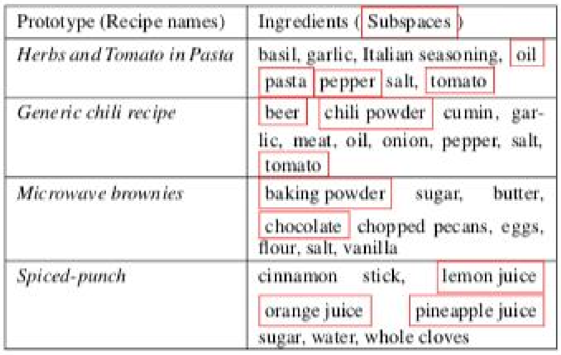
The dataset contains 56 recipes. Each recipe has 147 binary attributes. Each attribute represents the use of an ingredient. Despite each recipe is given a name, it is not used in this experiment. We want to identify the most important ingredients of each type of recipe. At first glance, there are two distinct (but related) tasks must be performed: putting the recipes into clusters and find the key ingredients in each cluster. We used the algorithm outlined in the previous section to create the indiscernible clusters. Over a million of indiscernible clusters were generated. The number of attributes (r) ranges between 1 and 22. Some of the representative results are reported below.

Table 2: A partial list of indiscernible clusters

r	Indiscernible Attributes	Count
2	baking powder, chocolate	2
3	beer, chili powder, tomato	2
3	lemon juice, orange juice, pineapple juice	2
4	oil, pepper, tomato, pasta	2

In the work reported in [2], the recipes were clustered into 4 groups. Bayesian Case Model was used to identify the key ingredients in each group (subspace) and are shown in the table below.

Table 3: Key ingredients identified using BCM. Reproduced from [2]



Prototype (Recipe names)	Ingredients (Subspaces)
Herbs and Tomato in Pasta	basil, garlic, Italian seasoning, oil pasta pepper salt, tomato
Generic chili recipe	beer chili powder cumin, garlic, meat, oil, onion, pepper, salt, tomato
Microwave brownies	baking powder sugar, butter, chocolate chopped pecans, eggs, flour, salt, vanilla
Spiced-punch	cinnamon stick, lemon juice orange juice pineapple juice sugar, water, whole cloves

VI. CONCLUSION

The preliminary results show that the key ingredients for the four groups of recipes can also be identified using discernibility relationship.

The current algorithm is, in many aspects, still very rudimentary and limited. It generates a huge number of indiscernible clusters, which is very difficult for humans to provide feedback. Some form of parametric control is required. For instance, it will be useful if a minimum frequency of attributes can be specified in indiscernible clusters. Future direction of this work will also include more robust experiments that involve more humans feedback.

REFERENCES

[1] Endert, A.& Fox, S. & Maiti, D.& Leman, S. & North, C. (2012). "*The semantics of clustering: analysis of user-generated spatializations of text documents*". AVI.

[2] Kim, Been & Rudin, Cynthia & Shah, Julie (2014). "*The Bayesian case model: a generative approach for case-based reasoning and prototype classification*". In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14), Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.), Vol. 2. MIT Press, Cambridge, MA, USA.

[3] So, R. (2018). "*Human-in-the-Loop Data Clustering - An Application in Higher Education*", International Conference on Engineering, Technology and Applied Science - Summer Session (ICETA), Hokkaido, Japan.

[4] Upadhyaya, Shuchita & Arora, Alka & Jain, Rajni. (2015). "*Rough Set Theory: Approach for Similarity Measure in Cluster Analysis*".

[5] Wenskovitch, J.E., & North, C. (2017). "*Observation-Level Interaction with Clustering and Dimension Reduction Algorithms*". HILDA@SIGMOD.

[6] Zhao, Yan & Yao, Yiyu & Luo, Feng. (2007). "*Data analysis based on discernibility and indiscernibility*". Information Sciences. 177. 4959-4976. 10.1016/j.ins.2007.06.031.