Topics in Algorithmic Data Analysis
Assignment 3: Explain It Like I'm Five

Muhammad Yaseen - 2577833
s8muyase@stud.uni-saarland.de

1 Introduction

In this report we consider three algorithms: DiffNorm [1], MGM [2], Desc + Disc [3] which we jointly refer to as EDD. These were proposed in the context of explanability and interpretability of user provided or algorithmically produced data clusters. In the following sections we compare these methods and look at their similarities and differences.

2 DiffNorm vis-à-vis MGM

DiffNorm operates on a set of databases and attempts to characterize the general distribution across the whole set (**norm**), as well as identify the locally important differences in the individual databases as well (**diff**). It can be used to characterize any arbitrary user-picked subset of datasets generated on the same alphabet. MGM takes as input a single dataset with some kind of category or class labels and produces an interpretable set of distinguishing features (called groups) which can explain or justify the given labels. The difference between a set of databases and a single database is more cosmetic than fundamental as we can join all the databases in the set and distinguish each record with a database id which simultaneously also acts as the class or category label. To model data DiffNorm uses an MDL based approach and finds a non-redundant lossless summary of the given data whereas MGM is based on learning and inference on a probabilistic graphical model.

DiffNorm can be considered hierarchical in the following restricted sense. Consider a set of datasets $\mathcal{J} = \{D_1, D_2\}$ and user-specified set of interest $\mathcal{U} = \{\{D_1\}, \{D_2\}, \mathcal{J}\}$. Thus we can find patterns specific to D_1, D_2 and then those common between D_1 and D_2. But the lowest granularity which is considered is a single dataset and thus we do not capture intra-dataset hierarchical structure only inter-dataset heirarchy.

In MGM dimensions are first consolidated into groups i.e. a **group** is a set of dimensions joined together with a logical connective (e.g. and,or). This can be likened to patterns used to cover a transaction t specially when the connective is "and". Like the requirement of non-overlap in patterns used for cover in DiffNorm, a dimension can only belong to one group i.e. dimensions don't overlap between groups and multiple such groups can be present in a transaction. Equivalently said, an observation can be said to be covered by the associated groups present in that observation. Unlike cover however, the whole of transaction need not be covered. Furthermore each dimension starts out in its own group like a singleton itemset. Number of possible groups is upper bounded by number of dimensions, whereas in DiffNorm any of the possible 2^I sets can be used in the coding set and hence used to cover the transaction. The number of clusters k remains a user provided parameter in MGM and has to be selected based on several runs and looking at some quality parameter e.g. mutual information between labels where available or ELBO. In DiffNorm it is provided by set $\mathcal{U}$

Lastly for MGM, each observation n belongs to exactly one latent cluster k. But it is important to note that it does not mean a 1-to-1 mapping between a set of group and cluster as a group can be present in more than one cluster. So in that sense fairly common group across all clusters can be said to capture the **norm** and a very exclusive group which is present in only a few or single cluster can be considered the **diff**. Thus not strictly hierarchical (as a cluster is not completely contained within another cluster) some shared groups can be identified to characterize a more coarse grained concept which may not correspond directly to a cluster.

3 Clustering or Classification?

To differentiate between clustering and classification we ask: Given an unmarked transaction t is there a way to assign it to a cluster in DiffNorm and MGM? One possibility is to assign it to the cluster whose associate group formulas are satisfied the most by it e.g. as used to label the clusters using majority class labels in the Recipes dataset example [2]. Similarly, for DiffNorm it can be assigned to dataset whose associated pattern set covers it the best in terms of MDL. This however is more like classification than clustering as we only assign it to the clusters we were given and do not produce any clustering from the algorithm itself.

Building on above, we can say that given only unlabelled or uncategorized data, neither of these methods will help in producing an understandable unsupervised clustering. Thus the methods are more useful as tools for explainability and interpretability of clusters than they are for producing unsupervised clustering. They can be combined with traditional clustering methods so that the clustering method is run first and then on the resulting clusters either MGM or DiffNorm is run to see if there is any interesting interpretation to the clusters. Hence, these methods should be considered as classification explanation methods rather than unsupervised clustering methods.

4 A Desc and a Disc

Following the spirit of DiffNorm, EDD aims to explicitly characterise the database partitions or components via **pattern components** in an interpretable manner such that both the similarities (norm) and differences (diff) can be explained. Data components or partitions mean the assignment of dataset rows to a particular cluster, whereas pattern components are the patterns which describe those clusters or partitions the best i.e. each pattern component consists of those patterns that are most representative for its associated component. Interpretability here means that the pattern sets related to each partition can be examined and thus both common and discriminating patterns can be identified.

EDD comprises of an alternating two step process. First DESC identifies the pattern components for an already given decomposition. Secondly, DISC takes these pattern components and tries to find a better decomposition. The quality of decomposition is assessed by how good the assigned pattern components are in modeling the pattern distribution in the respective component. If the pattern components perfectly describe the data partition i.e. all transactions in that component would be covered by the associated pattern set there would be no discrepancy in predicted and expected frequency of an itemset.

One major difference in EDD, compared to MGM and DiffNorm is that it does not require any category labels at all however if they are available it can be used to initialize the partitioning in **DESC**. Thus we are not limited to only inter-category structure like MGM and DiffNorm but rather can also capture more fine grained intra-category structure.

5 Commonalities between EDD and MGM

In EDD a data region is considered to be a distinct component if it has significantly different pattern distribution compared to other components, where the distribution is characterized by pattern sets and "significantly different" is measured by a point-wise approximation to KL-divergence measure w.r.t to the presence and absence of the pattern in component model S_j. This is similar to the distinguishing groups in MGM where such groups have different values across clusters i.e. a gap. Like MGM, a single pattern can be informative or important for multiple components. Furthermore, the learned assignment matrix A in EDD is conceptually similar to the grid plots in MGM. The clear benefit in EDD is that it is learned directly and automatically as part of the algorithm.

6 Possible extensions to MGM

MGM doesn't penalize long groups e.g. groups which contain many dimensions connected with an **or**. These are very loose rules or explanations and are the least informative. Comparing this to MDL based approach, if a longer itemset is reported as informative for a cluster it must necessarily be the case that it is as a whole frequent, otherwise the encoding wouldn't be optimal. Similarly, there's no way to add the minimum support requirement for groups in MGM.

Additionally, MGM limits a dimension to be only associated with one group. The motivation for doing so however was not made clear. Consider two informative patterns {rice, tomato, tofu} which a vegetarian might buy and {rice, tomato, chicken} for a non-vegetarian. In MGM once {rice, tomato} have been associated with a group there's no way to use them for the other group. In similar vein, there might be conditional dependence between a set of dimensions given another dimension i.e. a dimension is usually absent in a cluster but only present when some other dimension is present. This would also be missed by MGM because of this restriction.

7 Who's the most (trust)worthy of them all?

In general, on the basis of practical utility EDD is the most useful as it is the only method we have considered which can work completely unsupervised and does not require a pre-partitioned database or selection of a number of clusters. Also keeping in mind the points addressed in previous section, if a dataset with category labels is available and user is interested in only inter-database information it would be better to use DiffNorm on it than MGM as DiffNorm rules are more informative (no loose or's), succinct (longer rules are penalized), and cover more types of interaction (overlap in dimensions is allowed) than MGM.

8 Limitations

The algorithms considered do not incorporate domain knowledge i.e. the dataset as a whole may be unlabelled but there might be some well known interactions and constraints on the involved items e.g. in natural science experiments. That said, it is rather straightforward to incorporate this in the case of DiffNorm and MGM, for example if there is a well known restriction that two particles cannot coexist together (e.g. particle, antiparticle pairs) we can add a restriction set $\mathcal{R}$ so that the generated candidate c is rejected if $c \in \mathcal{R}$. This provides an easy way to find clusters while respecting the known constraints or interaction limits. It can also be used to prevent partitioning solely based on known interactions as they are useless from data exploration point-of-view. There is no such transparent way to do so in MGM, and the groups have to be removed post-hoc.

Given a pattern set as explanation, all patterns are assumed to have the same importance (in MGM and EDD) and the same is true for all items within a single pattern (all three). This may not be true in some settings e.g. social sciences and medical diagnoses where the importance of some symptoms or social indicators might be more important than others and users will want to get a feature importance ranked list of patterns and items. Gain in objective function for a pattern can be used as a proxy for pattern level importance, as in DiffNorm.

References

[1] K. Budhathoki and J Vreeken. "The Difference and the Norm – Characterising Similarities and Differences between Databases". In: *In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD'15)* (2015).

[2] B. Kim, J.A. Shah, and F Doshi-Velez. “Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction”. In: *In Proceedings of the 28th conference on Advances in Neural Information Processing Systems (NeurIPS’15)* (2015), pp. 2260–2268.

[3] S. Dalleiger and J. Vreeken. “Explainable Data Decompositions”. In: *In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (2020).