The Missing Link? On the In-Between Instance Detection Task

Daniyal Kazempour*, Claudius Zelenka*, Peer Kröger
{dka,cze,pkr}@informatik.uni-kiel.de
Kiel University
Germany

ABSTRACT
This vision paper reevaluates the binary view of "inlier" versus "outlier" in data mining, proposing the concept of "in-between instances" (IBIs) as a new category. The term "in-between instances" denotes objects that serve as connectors between multiple clusters, sharing traits from two or more classes. This concept differentiates from existing probabilistic or fuzzy clustering models, which may assign objects to multiple classes but fail to explicitly recognize the unique role of IBIs. This study aims to explore the IBI task, examining the problem's characteristics, its connections to other research fields, and its potential benefits in various domains. It also proposes potential archetype predicates capable of identifying IBIs, and indicating which classes they connect, marking a significant departure from traditional data mining tasks.

1 INTRODUCTION
When it comes to data mining there exist certain well-established unsupervised machine learning tasks that yielded prolific outcomes in the form of publications, among them the most prominent members are clustering and outlier detection. While in clustering the task is to partition a dataset in such a way that similar objects are grouped together and dissimilar objects are far apart [10], outlier detection aims for a different focus, being devoted to the detection of objects that express a certain distance, expressed in instances that appear anomalous w.r.t. instances of other clusters. For both tasks (clustering and outlier detection) their purpose and validity are undoubted within the community which is supported by the wealth of literature (i.e. for the case of clustering it raised the question of why there are so many algorithms [8] that has been published over the previous decades. We provoke here in this work this binary view of "inlier" (member of a cluster) vs. "outlier" by the following statement:

There exists a special type of instances that are characterized by their property to connect clusters, by acting as a (semantic) conduit between groups.
While they may not be inliers, they may as well not be real 'outliers' in the sense of being an anomaly.
They share certain traits from two or more neighbouring classes, acting as potential connectors.

We denote this novel special type of instance with the term 'in-between instance' (short: IBI). A visualization of this setting is shown in Figure 1 with two clusters of orange and green points and the IBI data point in red. While it may be argued that similar approaches that determine objects statistically belonging to multiple classes would be covered by probabilistic models such as EM clustering or fuzzy clustering, we disagree with these solutions being regarded as IBI detection methods. In fact, they may provide a class-affiliating quantity for each object, yet they fail to explicitly detect and discern in-between instances from 'real' outliers that are characterized by not sharing any traits with two or more classes.

From the question of which fields of applications this concept can potentially serve we provide in the following some example use-cases and their corresponding semantics of an IBI.
From Table 1 we can get a glimpse of the vast and diverse fields and different types of data in which in-between patterns can be discovered, including the potential meanings and benefits of finding such instances.
While we have in Table 1 some examples of what an existing in-between instance can mean and which potential benefit it can bring, there remains the question of what one can state in the case where no in-between instances are detected. This absence of in-between instances can have several causes, such as (but not limited to): (a) There exist no instances that share common properties between two or more classes (b) are there reasons for why there is an absence of in-between instances? and (c) can a remedy to this absence be provided or is that not possible (i.e. due to physio-chemical limitations)? and finally (d) what are the potential benefits if such an in-between instance is found or created (i.e. the creation of a novel drug)? From the previous implications, one may observe that we shifted here from the question of "Is there an in-between instance?" to "Why is there no in-between instance? Which properties from which classes would it embody if existent? Is there a need/benefit for an in-between instance and what are the reasons for it?"

The vision that we want to convey here is to enter and take on the journey of elaborating on the IBI task itself, investigating the
2 RELATED WORK

In this section, we discuss the connections and differences of IBI to other subfields in data mining.

2.1 On the Cluster Detection Task

Clustering is the task of partitioning a given dataset in such a way that objects within the same cluster are as similar as possible while objects between different clusters are dissimilar. This task is a common method across different domains [30]. To achieve a partitioning of a given dataset several approaches have been developed relying on different underlying assumptions. In the case of the DBSCAN [7] algorithm the assumption is that clusters are dense and separated by sparse regions. In the case of \( k \)-means [19] it is assumed that the variance of clusters discovered for a predefined number of partitions \( k \) is minimal. The in-between instance detection acts here in an orthogonal way. While clustering has the goal to maximize the dissimilarities between two or more clusters, in-between instances, if located between partitions, decrease the dissimilarity since such instances exhibit partial similarities to their surrounding clusters.

2.2 On the Anomaly Detection Task

Anomaly detection is the task of finding the exception from the norm in a dataset [1]. It is relevant to many applications, e.g. medical diagnostics, where it can help radiologists identify abnormalities in medical scans [9]. Anomaly detection can also be seen as outlier detection or out of distribution detection [22]. The task of in-between instance detection has similarities with anomaly detection in so far that in both tasks out of distribution data is the target. However, the key difference is that anomaly detection does not make assumptions about the location or direction of the out of distribution data (between two clusters) and only considers its distance to the closest in-distribution as in [20] using metric learning.

2.3 Soft Labels, Noisy Labels, and Regression

In a supervised classification labeling typically hard labels are used. This means that each datapoint is assigned a distinct label. For a binary classification problem, these labels can be encoded as 0 or 1. In contrast, soft labeling [29] describes the process of assigning labels that allow the expression of nuances, which for a binary classification means allowing all values between 0 and 1 as labels. This can be interpreted as a regression problem or signal ambiguity in the labels [26].

We want to clarify that soft labels or regression and in-between data points are two distinct concepts even though the IBI task requires a continuous output space. Just because a datapoint is assigned a soft label, even in high-dimensional spaces, does not mean that it is automatically an in-between instance. While real-world labels in supervised machine learning may suffer from random and human annotator-dependent uncertainty with changing variance and condition, i.e. aleatoric heteroscedastic uncertainty [25]. These noisy labels are not the target application of IBI detection.

3 DEFINITION OF IN-BETWEEN Instances

At first glance, it is tempting to consider in-between instances to be simply outliers. And in fact they would match the definition of outliers as stated by Hawkins [11]:

"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism."

In-between instances are considered to deviate indeed from any regular patterns e.g. distributions of clusters. Being aware of the diverse landscape of outlier models as elaborated by Zimek et. al [31], we refer in this work to the deviation-based outlier model that is described as follows:

"Deviation-based outlier detection groups objects captures some characteristics of the group, and considers those objects outliers that deviate considerably from the general characteristics of the group."

The question that arises at this point is: what is an in-between instance and as a consequence, what makes it different from outliers? For this, we provide first a definition of in-between instances:

**Definition 1 (In-between instance (IBI))**

Given a dataset \( X \in \mathbb{R}^{n \times d} \). Furthermore given a partitioning...
Given a dataset \( X \in \mathbb{R}^{n \times d} \). Furthermore, given a partitioning \( C = \{C_1, ..., C_k\} \subseteq X \) and a set of outliers \( O \subseteq X \) where \( C \cap O = \emptyset \) and \( \{X\} = C \cup O \). An object \( o \in O \) is an in-between instance if it satisfies a predicate \( \theta(o, C_i, C_j) \in \{\text{true, false}\} \) with \( i \neq j \).

According to definition 2, for a given set of partitions \( C \) and a set of outliers \( O \), an object \( o \) is considered as an in-between instance if it is an element of the set of outliers \( O \) and satisfies a predicate \( \theta \). The set of partitions \( C \) and the set of outliers \( O \) can either be obtained through labeling via clustering and outlier detection methods or through manual labeling of the data by domain experts.

The more crucial aspect lies in the meaning of the predicate \( \theta \) that has to capture the semantic of being in-between two or more clusters. The \( \theta \) functions are predicates that reflect if a potential in-between object \( o \) exhibits a certain similarity or proximity to at least two (or more) clusters. This similarity can be expressed e.g. in terms of (1) neighborhood (2) probabilities or (3) geometric orientation. A part of this vision is to propose which aspects can be relied on in terms of predicates to capture proximity and hence to model in-betweenness. In the following, we propose potential ways of representing these criteria \( \theta \):

1. **Neighborhood Criterion**: Given two objects \( x \in C_i, y \in C_j \) with \( i \neq j \) and an object \( o \in O \). Furthermore, we define \( N(x) \) as the neighborhood of an object (set of objects). \( o \) can be considered as in-between if the following predicate is satisfied:
   \[
   \theta(o, x, y) = o \in N(x) \wedge o \in N(y)
   \]
   where the neighborhood predicate relies on \( \varepsilon \)-range or \( k \)-nearest neighbors. The intuition behind this predicate is that an object is considered an in-between instance if it is in a certain neighborhood of objects from two (or more) different clusters as is illustrated in Figure 2.

   To rely on a neighborhood predicate is per se not new and has been successfully utilized in prominent clustering algorithms like DBSCAN [7] (and variants) as well as in outlier algorithms like LOF [2]. The application of a neighborhood-based predicate is, in particular, useful in spatio-temporal context where the concept of locality plays an important role as it has been demonstrated in case of epidemiologic monitoring, like tracking the spreads of the west nile virus [4]. The question of which range of neighborhood proximity and hence to model in-betweenness. In the following, we propose potential ways of representing these criteria \( \theta \):

2. **Class Probability Criterion**: Given two clusters \( C_i, C_j \) with \( i \neq j \) and an object \( o \in O \). Furthermore given for \( C_i \) and \( C_j \) their respective probabilistic model \( M_i \) and \( M_j \) (i.e. Gaussian distribution modeled through mean and covariance). \( o \) is considered as an in-between instance if the following predicate is satisfied:
   \[
   \theta(o, M_i, M_j) = p(o|M_i) \approx p(o|M_j)
   \]
   where \( p(x|M) \) denotes the probability of an object \( x \) belonging to a model \( M \). With this probabilistic-driven predicate, an object is considered as an in-between instance if it has similarly lower probabilities belonging to two (or more) clusters as illustrated in Figure 3 where the potential in-between instance is located in the overlapping contours from \( M_i \) and \( M_j \).

   The probabilistic concept is also used in prominent algorithms like expectation maximization (EM) [6] based clustering methods as well as in outlier detection such as in LoOP [16]. In LoOP the authors state that due to the nature of probabilities the obtained
outlier score is in a range between $[0, 1]$ facilitating the interpretation of the result for the users, especially for those that are not familiar with the theoretical foundation of the algorithm. Especially in the field of medical imaging probabilistic approaches like EM have been applied over the past decades [14].

(3) Geometric Property Criterion: Given two objects $x \in C_i$, $y \in C_j$ with $i \neq j$ and an object $o \in O$. Furthermore, we denote $d_{\cos}(x, y) = 1 - \frac{(x, y)}{\|x\|\|y\|}$ as the cosine distance between two objects. $o$ is considered as an in-between instance if the following predicate is met:

$$\theta(o, x, y) = d_{\cos}(x, o) \leq d_{\cos}(x, y) \wedge d_{\cos}(y, o) \leq d_{\cos}(x, y)$$

This angle-based predicate states that an object can be considered as an in-between instance if the angle between an in-between instance $o$ and an object $x \in C_i$ is smaller or equal to the angle between an object $x$ and $y \in C_j$ and likewise for $o$ and $y$ as shown in Figure 4.

Relying on cosine similarity and thus on the enclosed angle between vectors is popular in cases of information retrieval and text mining where due to the high-dimensional nature of the vectors cosine similarity is used as stated by Singhal [28]. Also, in the context of outlier detection in high-dimensional settings, it is relied on angle-based concepts as stated by Kriegel et. al [17] where they present their angle-based outlier detection method ABOD.

4 POTENTIAL BENEFITS AND CHALLENGES

Aside from the bold statement that the third type of instance (besides inlier and outlier) would be necessary, the reader is probably now asking: why should one care or actually want to detect in-between instances? We wish to highlight the potential benefits of the IBI instance task with an image labeling/classification problem.

The CIFAR-100 data set [18] is a standard dataset for the evaluation of classification algorithms [27] and provides a wide range of small images of diverse classes. For our experiment, we select all instances of images labeled as ‘Man’ and ‘Elephant’. For each image we calculate embeddings using a pretrained visual transformer in B/32 configuration of OpenAI CLIP [23]. Afterward, we apply UMAP [21] dimensionality reduction and see that as shown in Figure 5 the images of ‘Man’ and ‘Elephant’ form two distinct clusters. This is expected behaviour and confirms that the classes ‘Man’ and ‘Elephant’ represent distinct semantic concepts.

However, there may also be the case of a man dressed up in an elephant costume. While not part of the CIFAR-100 dataset, this image represents an ‘in-between instance’ of a man and an elephant. We visualize the CLIP embeddings of an example image with the same UMAP reduction in Figure 5 and where the red point shows the location of the IBI image.

A potential challenge is that ‘in-between’-ness even defined using nearest neighbor criterion requires a concept of direction. We are aware that embeddings such as CLIP are only trained on similarity and UMAP and other dimensionality reduction algorithms are also not designed to preserve that, nevertheless, we argue that at least some global structure is preserved, even though UMAP has a stochastic component.

To guarantee this property embeddings trained using metric learning would be ideal, as loss functions for metric, such as triplet loss in [12] explicitly enforce consistent distances. However, we argue that even embeddings not trained with metric loss can at least locally preserve distance and directionality if their representations capture semantic concepts. In Radford et. al. [24] we can see for example that unsupervised representation learning allows the affine interpolation in the latent space for more or less happy face generation.

An open question is whether to apply the in-between predicate before or after dimensionality reduction. Also, distances and directions are not preserved by all dimensionality reduction algorithms. We argue that globally consistent dimensionality reduction such as PCA [13], ICA [5] or variations give more guarantees, however we see in our example that it can work with UMAP dimensionality reduction, which gives mostly locally consistent clusters dependent on the parameters (we used the default parameters of the implementation [1] by the UMAP authors).

Another assumption is that the concept of IBI data requires that the clusters are not overlapping.

1https://github.com/lmcinnes/umap
In the previous section, we demonstrated the usefulness of the in-between concept in image classification and discussed possible challenges. Now we want to demonstrate potential real-world benefits with a biochemical application.

Obviously, such cases cannot appear in binary case or categorical one-dimensional datasets.

Another example application is in the context of drug discovery. While it may be well known to medical chemists which potential drug compounds are similar to each other, forming clusters, there are cases in which there is a special interest to discover compounds that are located in-between of two or more classes of drugs. The reason for that particular interest is that such molecules exhibit characteristics not only from one class of drugs but from two or more which may be desired in the context of drug design [15]. In Fig. 6 we can see a compound (IBI) that is located between two classes of molecules ($C_1$, $C_2$). For the case of $C_1$ we have pharmaceuticals with calming effects such as tranquilizers like Promazine ($C_1$, first row, first compound), medications for treating acute psychosis and bipolar disorders like Chlorpromazine ($C_1$, second row, first compound), or antidepressants like Desipramine ($C_1$, first row, last compound). In the case of $C_2$ we have pharmaceuticals that are used in the context of cancer treatment and autoimmune diseases such as 6-Mercaptopurine ($C_2$, second row, last compound) that is especially used against certain cases of leukemia, or the cytostatic medication 6-Thioguanin ($C_2$, second row, first compound). Here 6-Thioguanine is a so-called antimetabolite meaning that it disrupts metabolism at a cellular level and thus the cytokinesis (reproduction of cells). Similarly in its effect is Zoxazolamine ($C_2$, first row, first compound) that was originally used as a muscle relaxant in the early 1950s but was discovered to exhibit precisely such antimitabolite effects causing liver damage. Lastly, we have in $C_2$ a virostatic compound that suppresses the multiplication of virus-infected cells such as Favipiravir ($C_2$, first row, last compound).

As for the IBI named Phenazopyridine, one can observe common structural properties from both $C_1$ and $C_2$, such as the connection of two aromatic rings via a nitrogen atom (Figure 3, yellow rectangles) or the nitrogen-rich substructures (Figure 3, blue rectangles). The IBI comprises molecular substructures that are unique to either of the two classes. At this point, the questions emerge: does a composition of structural properties also follow a composition of pharmacutic properties, i.e. can this medication be used for both use cases (mental calming effects, and antibiotic/virostatic/cytostatic effects)? According to [3] in 2007 it was discovered that our detected IBI compound Phenazopyridine exhibits besides its analgesic (pain relieving) effect for the urinal tract also antibacterial properties that have been successfully used to cure puerperal fever.

In conclusion, through this simple and small example, we have an illustration that with research in the field of in-between instance detection scientists may also benefit in the domain of drug discovery.

5 EXAMPLE APPLICATIONS

In this paper, we present the vision of in-between instances and want to offer a fresh perspective besides the established concepts of clusters and outliers. We provide possible criteria and visualize the concept on CIFAR-100 and drug discovery tasks. Also, we briefly discuss both the potential and the challenges of the concept. Our hope is to get a broader discussion about IBIs started and inspire researchers in all fields to think anew about the potential insights from their data.

ACKNOWLEDGEMENT

We would like to thank at this point Melanie Oelker for her elaboration on the potential interest of in-between instance detection for medical chemists in the context of drug discovery.

REFERENCES